Université
de Toulouse

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Spécialité : **Informatique**

Délivré par l'Université Toulouse III - Paul Sabatier

Par

Bachelin RALALASON

Représentation multi-facette des documents pour leur accès sémantique

Présentée et soutenue le **jeudi 30 septembre 2010** devant le jury composé de **:**

Pr. Florence Sèdes	Université Paul Sabatier	Présidente du Jury
Pr. Cécile Roisin	Université de Grenoble	Rapporteure
Pr. Francis Rousseaux	Université de Reims	Rapporteur
Pr. Javier Pereira	Universidad Diego Portales, Santiago, Chile	Examinateur
Pr. Josiane Mothe	IUFM Midi-Pyrénées	Directrice de thèse
Mcf. Gilles Hubert	Université Paul Sabatier	Co-encadrant

INSTITUT DE RECHERCHE EN INFORMATIQUE DE TOULOUSE
Centre National de Recherche Scientifique (UMR 5505) - Université Paul Sabatier
118, route de Narbonne, 31062 Toulouse cedex 04, Tel. 05.61.55.66.11

Représentation multi-facette des documents pour leur accès sémantique

REMERCIEMENTS

Je tiens à remercier chaleureusement :

Monsieur Louis FARIÑAS DEL CERRO, Directeur de l'Institut de recherche en Informatique de Toulouse de m'avoir accueilli au sein du Laboratoire IRIT.

Monsieur René CAUBET qui m'a donné la chance de pouvoir préparer cette thèse et d'accéder au monde de la recherche grâce à la collaboration qu'il a mise en place entre l'Université Paul Sabatier et l'Université de Fianarantsoa. Dans cette collaboration, je remercie également M. Benjamin RANDRIANIRINA, président de l'Université de Fianarantsoa et M. Bertin RAMAMONJISOA, directeur de l'ENI.

Monsieur Claude CHRISMENT, Responsable de l'équipe Systèmes d'Informations Généralisés du laboratoire IRIT, pour son hospitalité irréprochable au sein de son équipe et tous ses précieux conseils et recommandations.

J'adresse mes vives reconnaissances à Madame Josiane MOTHE pour la bonne direction de cette thèse ; pour sa disponibilité, son encadrement, ses conseils et son soutien inestimable. Qu'elle soit ici assurée de mon très grand respect et du plaisir que j'ai à travailler avec elle.

Je tiens à remercier profondément à Monsieur Gilles HUBERT, pour son encadrement, ses conseils précieux et les fortes collaborations.

Je tiens également à remercier Mme Florences SEDES pour ses consignes et conseils qu'elle m'a donnés tout au long de mon séjour à l'IRIT et pour la participation à ce jury de thèse.

Je souhaite exprimer toutes mes reconnaissances à Madame Cécile ROISIN et à Monsieur Francis ROUSSEAUX pour l'honneur qu'ils me font en acceptant d'être les rapporteurs de ce mémoire.

Je remercie vivement Monsieur Javier PEREIRA qui a bien voulu accepter de faire partie de mon jury.

J'exprime mes profondes gratitudes à Mme Nadine Baptiste JESSEL ainsi que tous les membres du projet européen *Contrapunctus* pour les coopérations fructueuses que nous avons eues pendant la réalisation de ce projet.

Je souhaite exprimer mes sincères remerciements à Messieurs Bernard DOUSSET et Mohand BOUGHANEM pour leurs conseils et conversations instructives qui m'ont permis d'avancer d'avantage dans mes activités de recherche.

Je remercie également tous les membres du projet *Dynamo* pour leur cordiale collaboration.

Mes remerciements vont également vers tous les membres de l'équipe SIG pour leur convivialité durant mon séjour au sin du laboratoire.

Je remercie chaleureusement mes collègues Damien DUDOGNON et Joël MARCO qui ont participé à la partie évaluation des fonctions de similarité de ce travail.

Patricia STOLF et Nathalie HERNANDEZ pour leurs concours, critiques appréciables et précieux conseils ainsi que leurs rigueurs d'encadrement pendant le stage de DEA.

Une pensée particulière est adressée à mes collègues que j'ai côtoyés quotidiennement, Désiré KOMPAORE, Benoît ENCELLE, Estella ANNONI, Bouchra SOUKKARIEH, Dana KUNKUN et Joëlson RANDRIAMPARANY. Je leur exprime ma gratitude pour leur sympathie, soutien, encouragement et franche collaboration.

Je remercie Sébastien LABORIE pour les concertations fructueuses qui m'ont permis d'avoir un champ de vision plus large dans les domaines en relation avec mes thématiques de recherches.

Je remercie également Mme Christine FAVRE et Mme Annie MARSHALL pour leur aide, leur soutien et conseils pragmatiques pour surmonter les problèmes rencontrés lors de mon séjour en France.

J'aimerais profiter de l'occasion pour remercier sincèrement Chantal MORAND, la plus malgache des toulousaines, pour tous ses soutiens et conseils pratiques dans la vie courante à Toulouse.

Je tiens à remercier chaleureusement tous les personnels de l'IRIT sans exception.

J'exprime mes grands remerciements à l'église catholique malgache de Toulouse (FKMT), à l'église protestante malgache de Toulouse (FPMA), au groupe Ainga Gospel Community (AGC), à tous les amis musiciens ainsi qu'à la communauté malgache de Toulouse pour les bons moments que nous avons passés ensemble.

Je remercie également tous ceux qui, de près ou de loin, ont contribué à la réalisation de cette thèse.

Je remercie ceux qui ont veillé sur moi depuis toujours, ceux qui m'ont fait confiance, qui m'ont soutenu sans faille dans tous mes projets et qui ont accepté mes choix sans pour autant toujours forcément les comprendre, merci à tous les membres de la famille.

Je tiens à remercier profondément ma mère et mon regretté père qui ont toujours été d'un soutien inconditionnel, qu'ils soient ici en partie récompensés pour tout ce qu'ils m'ont donné.

Enfin, un immense merci à celle qui partage ma vie, Lackouly, elle qui a réussi l'exploit de toujours m'encourager à finir cette thèse et d'élever seule les enfants pendant le déroulement de cette thèse. Ces dernières années ont dû lui paraître bien longues loin du foyer familial qui m'a tant manqué.

Merci à tous et à toutes.

RÉSUMÉ

La recherche d'information est un domaine prépondérant dans la vie contemporaine car elle permet à tout un chacun de disposer d'éléments d'informations qui permettent d'agir et de prendre des décisions face à une situation donnée. En effet, l'objectif d'une recherche d'information est de s'informer, d'apprendre de nouvelles notions que nous ne maîtrisons pas. La disponibilité et la pertinence de ces nouvelles informations ont donc une très grande influence sur la prise de décision.

La plupart des moteurs de recherche actuels utilisent des index regroupant les mots représentatifs des documents afin de permettre leur recherche ultérieure suite à une requête d'un utilisateur. Dans ces méthodes, les documents et les requêtes sont considérées comme des sacs de mots, non porteurs de sens. L'une des méthodes innovantes actuelles est l'indexation et la recherche sémantique des documents. Dans ce cas, il s'agit de mieux prendre en compte le besoin de l'utilisateur en considérant la sémantique des éléments textuels. Nous nous intéressons à cet aspect de la recherche d'information en considérant une indexation et une recherche à base d'ontologies de domaine. Les solutions que nous proposons visent à améliorer la pertinence des réponses d'un système par rapport au thème de la recherche. Le discours contenu dans un document ou dans une requête ne sont pas les seuls éléments à prendre en compte pour espérer mieux satisfaire l'utilisateur. D'autres éléments relatifs au contexte de sa recherche doivent aussi être pris en compte. La granularité des informations à restituer à l'utilisateur est un autre aspect. Par exemple, dans le cas de systèmes question/réponse, une réponse courte est attendue ; parfois un seul terme peut suffire à satisfaire le besoin de l'utilisateur. Permettre une recherche à différents niveaux de granularité implique que le modèle d'indexation et de recherche sous jacent intègre la prise en compte de la structure des documents. Les métadonnées associées aux documents doivent également pouvoir être intégrées dans le modèle de recherche d'information. Ces métadonnées permettent de qualifier les documents ; leur usage peut permettre de s'approcher au mieux des besoins de l'utilisateur. Enfin, le contexte de recherche implique également l'usage de l'information et les contraintes pouvant être liées à l'utilisateur (mal-voyant par exemple). Nous nous sommes intéressés à ces différents aspects et avons développé un méta-modèle de représentation multi-facette des documents en vue de leur accès sémantique. Dans notre modèle, le document est vu selon différentes dimensions dont la structure logique, la structure physique, la sémantique des contenus ainsi que leurs évolutions dans le temps. Chaque document est représenté et annoté selon les différentes facettes de représentation. La facette « contenu » est ici traitée de façon spécifique. Nous avons plus précisément étudié au travers de cette facette les mécanismes d'indexation et de recherche sémantique s'appuyant sur une ontologie de domaine de référence. Ces mécanismes s'appuient sur la notion de graphes de concepts produits par l'indexation. Nous avons également proposé des mesures de similarité sémantique entre concepts et une fonction de similarité entre les graphes d'annotation des requêtes et ceux des documents. L'aspect dynamique (modification des documents par exemple) a également été étudié relativement à la facette contenu des documents.

Pour évaluer et valider nos solutions, nous avons instancié ce modèle dans trois domaines distincts : l'apprentissage en ligne, la maintenance automobile et les partitions musicales Braille. Nous avons également évalué les résultats en termes de rappel/précision des fonctions que nous avons proposées et montré leur supériorité par rapport à l'état de l'art.

ABSTRACT

Information retrieval is part of our contemporary life because it helps us to find information which helps us in acting and decision making. Indeed, the objective of any information retrieval task is to learn new facts, new notions. Thus, the availability and relevance of the pieces of new information we access have a high influence on decisions we make.

Most of the current search engines use indexes composed of the representative words from the documents; these indexes allow their access when compared to users' queries. These techniques consider documents and queries as bags of words but not the discourse they contain. One of the new methods to face the understanding of user's needs is semantic indexing and retrieval. In this thesis, we consider semantic indexing when based on ontologies that gather the domain knowledge. Matching content is not the only aspects that interest the user when searching for information. We consider other aspects such as the granularity of the elements to retrieve, the meta-data that can be associated with contents and the context in which the search is made. We consider these different aspects and propose a generic model based on a multi-facet representation. The facet related to document content is deeply studied. We made proposition related to semantic retrieval based on graph concepts and suggested a family of concept similarity functions and a graph similarity function that allow to compare graph concepts from documents and from queries. The dynamic aspect of the document collection has also been studied.

To validate this model we considered three application domains: e-learning, automobile diagnostic and Braille musical scores. We also evaluate our semantic similarity functions in terms of recall and precision and show their effectiveness.

TABLE DES MATIÈRES :

> Celui qui rencontre le plus de succès dans la vie est celui qui est le mieux informé.
>
> [Benjamin Disraeli]

Chapitre 1. INTRODUCTION GENERALE

Les hommes sont dotés de la faculté de raisonner, de choisir, et de prendre une décision. Le bon sens ou la raison est donc la chose au monde la mieux partagée. Néanmoins, le résultat de cette faculté de raisonner reste inexploitable et non fiable s'ils ne disposent pas de tous les éléments de base qui permettent de prendre la bonne et meilleure décision. Selon (René Descartes, *Discours de la méthode*), la diversité des opinions ne vient pas que les uns sont plus raisonnables que les autres, mais seulement de ce que nous conduisons nos pensées par diverses voies, et ne considérons pas les mêmes choses.

Ainsi, devant un fait ou un problème quelconque, seuls ceux qui disposent de ressources d'information complètes et de bonne qualité sont capables de bien raisonner et de trouver la meilleure solution.

À l'époque et dans les lieux où le traitement automatique de l'information n'existe pas encore, la recherche d'information (RI) est déjà présente dans la vie quotidienne et est réalisée en se contentant d'utiliser les moyens empiriques comme l'utilisation de répertoire ou annuaire, d'index manuel ainsi que des annotations et résumés pour accéder aux documents contenant l'information recherchée.

À l'heure où la plupart des documents sont numériques voire se trouvent sur la toile, il est de rigueur de disposer des outils permettant de rechercher, parmi tant d'autres, les bons documents contenant les informations qu'un utilisateur donné tente de rechercher, pour prendre une décision ou bien pour enrichir ses connaissances.

Naturellement, les besoins en information varient d'une personne à l'autre selon les contextes et circonstances dans lesquelles celles-ci expriment leurs besoins. De même, la manière utilisée pour exprimer un même besoin d'information peut différer d'un utilisateur à l'autre, compte tenu de ses acquis et expériences personnels.

L'objectif des systèmes de recherche d'information (SRI) est donc d'une part de permettre aux utilisateurs d'exprimer facilement leurs besoins d'information au moyen des requêtes, et d'autre part de leur fournir les documents potentiellement pertinents par rapport aux besoins d'information qui ont été transmis au système.

La RI sur la toile ne peut se faire qu'à l'aide des moteurs de recherches ou SRI. Plusieurs types de SRI sont disponibles actuellement et chacun a ses spécificités. Les uns, qui parcourent et analysent en permanence la toile, n'ont aucune vision sémantique du contenu d'un site alors que d'autres, qui sont plutôt des portails avant d'être un moteur de recherche, recensent par le travail de documentalistes, des sites et les différents sujets ou rubriques qu'ils traitent. Les méta-moteurs de recherche quant à eux sont des outils qui effectuent leurs recherches à partir de résultats donnés par d'autres moteurs de recherche et qui en résument les résultats. De même l'accès à la documentation que manipule une entreprise de façon interne doit également reposer sur des SRI. Tous ces types de système ont en commun la capacité de retrouver les documents potentiellement pertinents et de les proposer aux

utilisateurs au moment de la RI. Tous les documents doivent être préalablement indexés par le système. Les index associent aux documents les termes significatifs qui permettent non seulement de représenter les contenus des documents mais également de les retrouver ultérieurement par comparaison entre les termes de la requête et ceux des documents.

Les moteurs actuels ne prennent pas suffisamment en compte le contexte de la recherche et c'est cette problématique générale que nous étudions dans cette thèse à travers différentes facettes.

1.1. Indexation et recherche sémantique

La recherche de documents relatifs à un thème donné est une activité fréquente qui est indispensable pour compléter ou affirmer des connaissances préalables mais également pour en acquérir de nouvelles. À l'ère de l'utilisation massive des documents papiers, les documentalistes ont utilisé des méthodes manuelles afin de pouvoir retrouver facilement les documents que les lecteurs veulent consulter ou emprunter. Parmi ces méthodes, nous pouvons citer l'indexation analytique qui se traduit par l'utilisation de descripteurs (mots-clés issus d'un thésaurus) pour représenter les contenus, des classifications systématiques par thème ou champ disciplinaire ainsi que l'utilisation de notices qui qualifie le document au travers de métadonnées et d'un résumé.

Cependant, pour un grand nombre de documents, cette méthode manuelle n'est plus applicable car elle est très couteuse en temps de traitement. Une amélioration et automatisation de cette méthode s'imposait donc afin de réduire le temps de traitement de chaque document.

Dans les années 70, des méthodes d'analyse de textes ont ainsi été développées afin d'en extraire les termes dits caractéristiques de chaque document. Cet ensemble de termes caractéristiques, appelé aussi index et construit automatiquement, pour chaque document a remplacé l'utilisation des descripteurs et des résumés qui sont associés manuellement à chaque document. L'aspect sémantique qui était implicite via l'utilisation des thésaurus dans la démarche manuelle a été substitué par une analyse syntaxique automatique, permettant de traiter les variantes des mots mais pas leur sens.

Dans cette approche automatique, chaque terme d'indexation n'est pas considéré comme ayant la même importance. Cette hypothèse a conduit à la pondération des termes représentatifs de chaque document par le biais de la méthode Tf*Idf qui indique l'importance relative de chaque terme pour chaque document (Sparck Jones, 1972).

La recherche de documents potentiellement pertinents pour l'utilisateur consiste ensuite à comparer les termes constituant les besoins en information ou requête exprimée par l'utilisateur par rapport aux termes représentatifs de chaque document.

Cette méthode qui est prépondérante dans les systèmes actuels n'est pas satisfaisante lorsque l'on considère les avancées dans le domaine de la linguistique et des technologies en lien avec la gestion des connaissances. Les avancées dans ce domaine laissent penser qu'il est possible de concevoir des systèmes qui ne doivent pas se contenter de retrouver les documents contenant les mots de la requête mais qu'ils doivent retrouver ceux répondant sémantiquement à la requête.

L'apparition de la RI sémantique est une réponse possible à ce nouveau challenge en RI. Cette façon de rechercher les documents exige que l'utilisateur et le système utilisent une même base de connaissance en matière de sens à accorder aux différents termes présents, aussi bien dans les documents que dans la requête utilisateur. Des recherches ont déjà été faites dans ce domaine. Parmi les solutions proposées se trouve principalement l'utilisation des ontologies

qui sont des représentations formelles et partagées des connaissances d'un domaine. Cette représentation de connaissances est censée faciliter la communication entre la machine et l'utilisateur. Ainsi, les ontologies, qui sont des ensembles de concepts et de relations entre ces concepts, ont été utilisées dans plusieurs travaux (Bouzeghoub et al., 2005), (Song et al., 2005), (Hernandez et al., 2006), (Chang et al., 2007) comme un vocabulaire de référence pour indexer les documents électroniques afin de prendre en compte la sémantique des documents de la collection. Cette indexation basée sur l'utilisation d'ontologies, communément appelée indexation sémantique, consiste à associer à chaque terme représentatif des documents des concepts de l'ontologie pour spécifier le sens accordé à ces termes.

L'utilisation d'une ontologie comme base d'indexation a engendré de nouvelles problématiques à plusieurs niveaux auxquels nous nous sommes intéressés.

Une des problématiques fondamentales liées à l'utilisation d'ontologies comme base de référence des termes d'indexation concerne la forme du résultat de l'indexation. Il peut être un ensemble de concepts isolés ou un ensemble de graphe de concepts, reliés entre eux par des liens sémantiques. Dans le premier cas, en termes de résultat de l'indexation, la différence entre l'indexation sémantique et l'indexation sac de mots concerne le type de termes d'indexation (concept à la place de mots ou groupe de mots). Dans le second cas, la nature même du résultat de l'indexation est modifiée ce qui implique de mettre en œuvre de nouveaux mécanismes permettant l'appariement entre les documents et les requêtes.

Un autre aspect concerne l'importance accordée aux concepts lors de l'indexation. En effet, l'annotation sémantique d'un document consiste à associer des concepts d'une ontologie à un document généralement via les termes contenus dans ce document. Les concepts associés aux termes du document ont chacun leur importance vis-à-vis du document. Une pondération reflétant non seulement la sémantique des concepts mais également l'importance relative entre concepts par rapport au contenu du document doit être mise en place. De plus, la sémantique associée à certains documents n'est pas forcément associée à un texte particulier du document mais surtout dégagée à partir du document entier.

Enfin, une autre problématique de l'indexation sémantique est la non disponibilité d'un vocabulaire unique d'indexation. Des ontologies décrivant les connaissances correspondant à des domaines spécifiques existent, mais aucune de ces ontologies ne peut être appliquée à d'autres domaines que celui pour lequel l'ontologie a été construite. Il est donc difficile d'indexer avec une seule ontologie. Les mécanismes d'indexation doivent donc être génériques pour pouvoir s'adapter aux ontologies rencontrées.

1.2. Documents : évolution et mise à jour des index

Le cycle de vie des documents, qui sont au centre de tout SRI, présente plusieurs problématiques que nous détaillons ci-après.

Dans le contexte actuel d'utilisation massive des documents électroniques, la collection de documents ne cesse d'évoluer car de nouveaux documents sont ajoutés et d'autres sont supprimés de la collection. Cette évolution qui peut survenir à fréquence très élevée engendre l'indisponibilité des documents car les index qui permettent l'accès aux documents doivent être remis à jour. Autrement, les nouveaux documents ne seront jamais retrouvés.

De la même manière, les modifications apportées aux documents déjà présents dans la collection doivent elles aussi faire l'objet de mise à jour des index pour que les modifications soient prises en compte. Effectivement, l'ajout, la suppression ou la modification de texte dans un document peut entrainer la suppression de concepts annotant le document, ou

nécessiter d'associer de nouveaux concepts aux textes nouvellement insérés dans le document.

Ces problèmes sont d'autant plus importants sur le web où la fréquence de mise à jour de la collection et des documents est très élevée. Par ailleurs, plus la fréquence de mise à jour de la collection est élevée, moins l'index est disponible pour la RI car il est à tout moment en cours de modification. En supplément, le temps de remise à jour des index est fortement lié au nombre de documents de la collection.

1.3. Contextes des recherches

Le contexte thématique d'une recherche est une des facettes centrale d'une recherche d'information. Cependant, elle n'est pas la seule.

La granularité des informations à restituer à l'utilisateur est un autre aspect. Par exemple, dans le cas de systèmes question/réponse, une réponse courte est attendue ; parfois un seul terme peut suffire à satisfaire le besoin de l'utilisateur. Permettre une recherche à différents niveaux de granularité implique que le modèle d'indexation et de recherche sous jacent intègre la prise en compte de la structure des documents.

Les métadonnées associées aux documents doivent également pouvoir être intégrées dans le modèle de recherche d'information. Ces métadonnées permettent de qualifier les documents ; leur usage peut permettre de s'approcher au mieux des besoins de l'utilisateur.

L'accessibilité au document et au contenu du document doit être prise en compte. Il s'agit ici de considérer le format des documents lors de la recherche. Les mal ou non voyants ont par exemple besoin des formats et de types de documents spécifiques pour qu'ils puissent accéder aux contenus des documents, même si les systèmes de recherche peuvent les trouver.

1.4. Contributions de la thèse

Dans cette thèse, nous nous sommes attachés à proposer un modèle de SRI permettant de prendre en compte les différents aspects évoqués plus haut. Nous avons ainsi axé notre recherche sur ces trois aspects suivants :

- la représentation multi-facette des documents. Chaque facette correspond à un des aspects utiles à la représentation des documents pour en permettre la recherche efficace et pertinente. Une facette commune à toutes les applications concerne la représentation des contenus. Les autres facettes peuvent dépendre des applications et peuvent concerner la représentation de la structure des documents, l'usage ou la tâche associée aux documents, etc.,
- l'indexation sémantique. Nous avons particulièrement étudié l'indexation à base d'ontologies des contenus qui permet l'accès sémantique aux documents. Notre modèle se base sur l'utilisation de graphes de concepts qui sont des représentations plus riches que de simples concepts isolés. Nous avons également fait des propositions relativement à la garantie de la disponibilité des documents et la cohérence entre les contenus des documents, les index de recherche et l'ontologie de domaine de référence pour prendre en compte la modification de la collection de documents,
- et enfin la recherche sémantique des documents. Dans ce cadre, nous avons été amenés à proposer une famille de fonctions de similarité entre concepts d'indexation et une mesure de similarité permettant de comparer les graphes de concepts représentant des documents et des requêtes.

Afin d'étudier la généricité du modèle, nous l'avons mis en œuvre dans différents domaines d'application.

- L'apprentissage en ligne. Toute activité de RI a pour objectif de trouver des documents contenant des informations importantes dont nous avons besoin pour nous informer ou pour approfondir une notion. Il nous a donc paru intéressant d'étudier le cas spécifique des systèmes d'apprentissage en ligne. Ils offrent un cadre et un environnement qui pourraient être transposés aux activités d'auto-apprentissage, caractéristique des recherches sur le web. Dans ce domaine, nos contributions portent essentiellement sur la ré-utilisabilité des objets pédagogiques, la modélisation des scénarii d'apprentissage et des méthodes pédagogiques et le suivi des connaissances supposées acquises par un utilisateur en fonction des contenus et les connaissances véhiculées dans les documents qu'il a consultés.
- L'aide au diagnostic automobile. Cette application est le résultat d'un projet national ANR auquel nous avons contribué (Dynamo 2008-11). C'est dans ce cadre que nous avons développé différentes fonctions de similarité entre concepts et une fonction de mise en correspondance entre les documents et les requêtes pour implanter une indexation et une recherche sémantique. Nous avons suivi l'hypothèse selon laquelle certaines applications nécessitent d'adapter les mesures de similarité sémantique par rapport à des besoins spécifiques. De fait, une mesure de similarité sémantique doit refléter les propriétés des rapprochements sémantiques entre les concepts d'une ontologie. Une autre contribution majeure dans ce cadre concerne l'évaluation de ces propositions dans un cadre réel.
- Les partitions musicales Braille. Nous nous sommes intéressés dans ce cadre à l'accessibilité des documents dans le contexte de mal ou non voyants. Notre contribution majeure dans ce cadre est la proposition d'un langage de balisage XML spécifique aux partitions Braille. Cette contribution a été intégrée au projet Européen Contrapunctus.

1.5. Plan de thèse

Les contributions que nous avons résumées ci-dessus sont développées dans ce mémoire selon l'organisation suivante :

Le chapitre 2 présente l'état de l'art dans les domaines traités dans cette thèse. Nous y présentons donc les travaux reliés concernant la modélisation des documents, les mécanismes généraux relatifs à la RI puis le cas spécifique de la RI sémantique et de la prise en compte de la dynamique en RI. Enfin, nous présentons les travaux dans le domaine de l'apprentissage en ligne.

Le chapitre 3 présente le modèle multi-facette de documents. Ce modèle générique permet de prendre en compte non seulement la facette relative au contenu sémantique du document mais également les autres facettes pouvant mieux qualifier le document. Nous déclinons ce modèle dans les différents cadres applicatifs que nous avons étudiés. Nous approfondissons ensuite l'étude de la facette relative au contenu des documents. Nous y étudions la représentation sémantique des documents via l'indexation sémantique à base d'ontologies ainsi que la dynamique des documents et nos mesures de similarités sémantique pour la recherche des documents.

Le chapitre 4 présente plus spécifiquement les développements et évaluations que nous avons conduits pour valider nos propositions. Nous présentons ainsi l'outil d'apprentissage en ligne

tenant compte de la gestion des connaissances des utilisateurs (PALM : Plateforme d'Apprentissage en Ligne Multimedia), la recherche sémantique de documents de maintenance automobile dans le cadre du projet Dynamo (Dynamic Ontology for Information Retrieval), ainsi que le langage de description des partitions musicales Braille BMML (Braille Music Markup Langage) développé dans le cadre du Projet Contrapunctus.

Enfin nous terminons, dans le chapitre 5, par des conclusions et perspectives sur nos travaux de recherche.

La folie, c'est se comporter de la même manière et s'attendre à un résultat différent.

[Albert Einstein]

Chapitre 2. ÉTAT DE L'ART

L'abondance des documents électroniques disponibles dans notre quotidien a nécessité l'automatisation de la recherche de documents afin de disposer le plus vite possible des documents qui correspondent mieux à nos attentes. Diverses méthodes ont été élaborées afin de permettre ces recherches automatiques de documents afin de satisfaire les besoins en information des utilisateurs en termes de qualité des résultats retournés et de délais de traitement. Dans ce chapitre, nous présentons les différents procédés utilisés jusqu'ici pour retourner les bons documents parmi ceux qui sont accessibles dans une collection de documents. Avant de détailler ces méthodes de recherche de documents, nous allons d'abord définir ce qu'est un document, puis analyser les modèles de représentation de documents en général et des documents musicaux en particulier. Ensuite, nous présentons successivement les méthodes de RI non sémantiques, les méthodes de RI sémantique, et la prise en compte de la dynamique et des contextes en RI. Enfin, nous exposons la RI dans le cas de l'apprentissage en ligne.

2.1. Modélisation des documents

2.1.1. Document : définition

Il existe de nombreuses définitions du document. L'origine latine de ce terme est «documentum» qui indique leçon, exemple, modèle et accessoirement preuve, texte.

Selon l'AFNOR, un document se définit comme « Ensemble d'un support d'information, quel qu'il soit, des données enregistrées sur ce support et de leur signification, servant à la consultation, l'étude, la preuve ou la trace, etc. : livre, échantillon de parfum, tissus, film, etc. Le tout constitue une unité autonome ». (AFNOR, 1987)

2.1.2. Modèle de document

Les systèmes complexes sont souvent abordés selon différentes dimensions dans lesquelles les objets d'étude peuvent être projetés, c'est-à-dire réduits en des formes simplifiées, sans perdre le sens relatif à leur étude (Stern, 1997). Selon (Pogodalla et Dury, 2004), il convient de distinguer les modèles conceptuels des modèles opérationnels : les premiers aident à comprendre, expliquer et analyser les problèmes tandis que les seconds offrent des moyens pratiques de résoudre les problèmes d'ingénierie ou d'implanter des architectures efficaces pour le traitement logiciel.

2.1.2.1. Modèles conceptuels de document

Selon (Stern, 1997) un document possède quatre dimensions :

1) la forme est « l'interface qui permettra à celui qui le consulte de prendre connaissance du contenu même du document ».

2) le contenu est « la somme des connaissances, "empaquetée" dans l'emballage » (le document est ici vu comme un emballage et son contenu), il s'agit de l'information mise en forme.

3) la structure est « la manière dont les différents éléments d'information contenus sont organisés, structurés, articulés entre eux »

4) l'identité concerne les informations sur le titre du document, son auteur et la date de sa réalisation.

(Pogodalla et Dury, 2004) ont identifié l'espace et le temps comme des dimensions physiques indissociables du processus de restitution, contenu et forme comme des dimensions logiques. De plus, ils mettent en exergue la nécessité de distinguer les dimensions intrinsèques et extrinsèques aux documents. Ainsi, l'espace intrinsèque permet de décrire l'organisation interne de la présentation d'un document (comme les positions relatives des figures et paragraphes dans une page), alors que l'espace extrinsèque englobe les opérations d'échange ou de diffusion de documents. Le temps intrinsèque quant à lui est utile pour la modélisation de flux vidéo synchronisés, tels que mis en œuvre dans les documents multimédias, alors que le temps extrinsèque est adapté, par exemple, à la modélisation des diverses phases de création de versions au sein de dossiers documentaires.

(Roisin, 1998) considère un document comme un ensemble des composants de base organisés suivant quatre manières de structuration. Ces niveaux de structuration peuvent être considérés comme quatre dimensions indépendantes dont :

- la dimension logique qui concerne l'organisation des informations en composants hiérarchiques (chapitres, sections, paragraphes, etc…),
- la dimension spatiale qui se rapporte à la mise en page des documents, leurs présentations à l'aide des feuilles de styles,
- la dimension hypermédia qui recouvre les liens hypertextes et les différentes actions. Cette dimension permet les différentes formes d'interaction sur le document présenté, comme la navigation intra- ou extra-documents par le biais d'hyperliens,
- la dimension temporelle qui se réfère aux synchronisations multimédia ainsi qu'aux descriptions des scénarios.

Cette façon de modéliser les documents permet de représenter de manière homogène la plupart des catégories de documents dont les rapports techniques, les lettres, les articles scientifiques ainsi que les documents avec des structures graphiques ou hypermédia. Ainsi, la définition d'un format de représentation des documents multimédia doit permettre d'exprimer les caractéristiques relatives à chacune des différentes dimensions de ces documents.

Pour ce qui est du temps, c'est un élément prépondérant dans la vie du document numérique (Roisin et Sèdes, 2004). Il y a plusieurs aspects du temps car il intervient dans le cycle de vie du document, sa durée de vie, la variabilité des contenus, la dynamicité des structures, les indicateurs linguistiques et la multiplication des versions. La prise en compte du temps implique de nouvelles approches pour la modélisation et l'indexation des documents car la relation entre temps et documents est complexe. Elle couvre les éléments suivants :

- le temps inhérent au média qui définit le temps de défilement des médias continus mais aussi leur synchronisation,
- le temps lié à l'évolution des documents, à la prise en compte de leurs versions,

- le temps dans la production de documents qui intervient via l'adaptation au profil utilisateur, à ses préoccupations ou comportement, au dispositif d'affichage à partir duquel il consulte,
- le rapport entre le temps de lecture et le temps du contenu, qui définit la validité ou l'obsolescence des informations au moment de leur restitution.

Les sciences informatiques, par l'outillage qu'elles développent, se focalisent plus sur l'étude du temps documentaire, constitué des relations temporelles entre les unités documentaires d'un même corpus ou à l'intérieur d'un même document.

(Maître et al., 2004) développe l'idée selon laquelle les documents XML intègrent une structure temporelle implicite : ordre des éléments, attributs calendaires ou horaires, indices textuels… L'objectif est de générer un document « temporalisé » afin de permettre l'exploitation.

2.1.2.2. Modèles opérationnels de document

Les modèles opérationnels sont par nature étroitement liés aux algorithmes de transformation documentaire (Pogodalla et Dury, 2004). Quand on parle de format de document, ceci sous entend implicitement une notion très concrète de modèle de document, correspondant ainsi avec des hypothèses fortes directement liées aux modalités de traitement de ces formats de données qui sont avant tout conçus pour le stockage et les échanges de documents.

La connaissance des différents formats de documents tels que, PDF, HTML, XML, RTF, GIF, TIFF permet de réaliser les différentes transformations possibles d'un document d'un format vers un autre, sans perdre les contenus. Dans beaucoup d'applications manipulant des documents dans différents formats, les transformations de documents sont une problématique clé. Chaque format possède ses particularités et ses utilités dans différentes applications. Par exemple, le format *PDF* d'Adobe, qui est construit autour d'un modèle général à base de pages, se focalise d'avantage sur la publication via un réseau électronique que son prédécesseur *postscript*, qui lui était délibérément orienté vers l'impression papier. *HTML* et *HyTime* sont spécialisés dans la navigation au sein de réseaux de documents connectés par des «hyperliens».

L'analyse des modèles opérationnels montre qu'ils sont liés aux objectifs initiaux de leurs concepteurs (édition numérique, microinformatique…) mais leur dynamique exprime une évolution vers une abstraction de forme croissante (Pogodalla et Dury, 2004). Plus précisément, le principe de séparation contenu/présentation ouvre un large champ aux transformations documentaires, dont une des caractéristiques est de favoriser les changements de forme.

2.1.3. Documents musicaux

2.1.3.1. Partitions musicales

Une partition musicale est un document qui contient toutes les informations musicales sur une composition musicale donnée. Les musiciens peuvent écrire ou lire une partition pour jouer et exécuter la musique écrite. En utilisant les partitions musicales, les musiciens communiquent, partagent, apprennent et composent de la musique. Selon (Rousseaux, 1990), une pièce musicale possède quatre représentations dont :

- la représentation gestuelle qui comprend la représentation dynamique et en temps,
- la représentation graphique qui est une représentation statique ou dynamique hors temps,
- la représentation auteur qui est un ensemble de descripteur permettant aux musiciens de décrire leurs œuvres,
- la représentation en EPF (Éléments Porteurs de Forme), qui en parallèle avec la représentation auteur, structure la base de données musicale et représente la dimension sémantique de l'œuvre.

Divers standards ont été définis pour aider le développement de langage qui permet de manipuler le contenu musical, leur représentation et leur relation.

Parmi les standards musicaux, SMDL (Standard Music Description Language) et SMR (Symbolic Music Representation) sont les plus importants. Nous donnons des informations plus détaillées sur ces standards dans la section suivante.

2.1.3.2. Standards musicaux

a) SMDL

Le SMDL est une norme qui sert à représenter l'information musicale que cette dernière soit seule ou combinée avec du texte et des graphiques. L'information multimédia est également supportée. Son objectif principal est de permettre l'échange de l'information musicale et multimédia exprimée selon une notation musicale commune et ce, dans l'optique d'être complet, flexible et facile à utiliser. Les différents formats musicaux existants respectent quelques orientations principales qui sont définies dans SMDL, à savoir :

- le domaine logique, ou cantus, est le contenu de base de l'information musicale,
- le domaine visuel décrit la typographie musicale, c'est-à-dire comment le domaine logique apparaît sur la partition au niveau des symboles, de la position des notes, des polices de caractères et de la mise en page,
- le domaine gestuel concerne l'exécution particulière d'un cantus. Il spécifie donc quand et comment les éléments du domaine logique doivent être interprétés,
- le domaine analytique consiste en des commentaires généraux et des analyses théoriques sur les informations contenues dans les trois autres domaines.

Un document SMDL peut donc contenir un cantus, un ou plusieurs liens vers les instances de ce cantus (fichier NIFF ou GIF (domaine visuel)) et un lien vers l'instance de la performance du cantus (fichier MIDI (domaine gestuel)).

b) SMR

Symbolic Music Representation (SMR) est un ensemble de recommandations pour représenter les informations musicales. Une représentation symbolique musicale est une structure logique basée sur les éléments symboliques représentants les éléments audiovisuels. Ces recommandations suggèrent la représentation de plusieurs aspects pour l'encodage des informations musicales dont le domaine orienté contenu et celui orienté présentation.

Concernant l'aspect présentation des informations musicales, même si la recommandation parle de l'accessibilité pour les mal voyants, elle ne définit pas comment implémenter cette accessibilité. SMR généralise le concept de notation musicale pour modéliser les aspects visuels d'une partition musicale, ainsi que les informations audio et annotations relatives à une pièce musicale.

En se basant sur ces standards des documents musicaux, divers formats ont été développés pour encoder les partitions musicales. Nous présentons dans la section suivante ces différents formats.

2.1.3.3. Formats d'encodage musicaux

Dans la littérature, plusieurs langages existent pour transcrire les informations musicales. Parmi eux, nous pouvons citer MusicXML qui est le format le plus utilisé pour les éditions musicales, la notation NIFF qui est un format orienté graphique pour l'échange d'information entre les programmes d'acquisition et les programmes d'édition des musiques en noir, et enfin MIDI qui mémorise les musiques en termes de notes à jouer et l'instrument avec lequel elles sont jouées.

a) MusicXML

MusicXML (Good , 2002) est un code XML permettant la description (note, rythme) des données musicales. Ce format est destiné à être interprété par tous les types de logiciels musicaux. Il est conçu pour être un format d'échange universel de la notation musicale, d'analyse, de recherche et d'exécution. Comme il est utilisé par beaucoup de programmes de notation, des séquenceurs, des programmes d'exécution musicale et des programmes d'enseignement, il est devenu le standard de facto dans le domaine musical.

b) NIFFML

NIFFML est une implémentation XML de NIFF (Notation Interchange File Format) (http://www.music-notation.info/en/niffml/niffml.html) qui est un format de fichier binaire conçu pour encoder de façon très précise les graphiques utilisés pour représenter les partitions musicales. Ainsi, il permet l'échange de données de notation musicale entre les logiciels d'édition, de publication et d'acquisition des partitions musicales. NIFFML permet donc d'étendre les fonctionnalités de NIFF.

c) MIDI

Musical Instrument Digital Interface (MIDI) (http://www.midi.org/about-midi/specshome.shtml) est un standard musical de communication entre les instruments et logiciels séquenceurs MIDI. En utilisant le protocole MIDI, ces derniers peuvent s'échanger

des données pour jouer, éditer et enregistrer des musiques. MIDI est un code conçu pour véhiculer des informations sur les sons. Par exemple, un clavier MIDI transmet seulement vers un générateur de son des informations et ordres tels que la hauteur des notes, l'octave, la vélocité, la durée et le nom du timbre à jouer.

d) Play Code

Play Code (http://www.dodiesis.com) vise à offrir une opportunité d'échange d'information entre musiciens voyants et mal voyants grâce au logiciel Braille Music Editor[1] (BME).

Ce langage a été développé spécialement pour coder les partitions musicales Braille. La plupart des spécificités de la musique Braille ont été pris en compte dans Play Code. Cependant, il s'agit d'un langage élaboré dans un format propriétaire. Ceci a empêché non seulement son évolution mais également son utilisation par d'autres outils Braille.

2.1.4. Conclusion

Les documents ont été modélisés d'une part par des modèles conceptuels de documents (Pogodalla, 2004), (Roisin, 1998) pour pouvoir comprendre et analyser leurs contenus et d'autre part par des modèles opérationnels pour solutionner les problèmes du traitement des documents. Néanmoins, il manque dans ces modèles d'une part la considération de l'utilisateur et du contexte dans lequel il veut utiliser le document et d'autre part la description sémantique du contenu des documents afin de mieux comprendre les messages véhiculés dans les documents.

Concernant les lacunes des formats d'encodage des partitions musicales, nous pouvons constater que malgré l'existence de différents formats d'encodage de partitions musicales qui sont issus des standards de représentation des formats musicaux, aucun d'entre eux ne supporte parfaitement la représentation et la manipulation des partitions Braille. Play Code convient bien pour décrire les partitions musicales Braille. Cependant, c'est un code propriétaire et ne peut pas être réutilisé ni étendu facilement. En outre, ni MusicXML, ni NIFFML, ni MIDI ne prennent en compte la notation musicale Braille. Ces raisons justifient le besoin de développer un nouveau format d'encodage des partitions musicales Braille.

Dans la section 2.2, nous présentons les différents composants de la recherche d'information dont l'indexation automatique de documents et les modèles de recherche.

2.2. Recherche d'information

L'objectif des SRI est de fournir aux utilisateurs les documents potentiellement pertinents par rapport aux besoins qu'ils expriment.

Les SRI utilisent des listes inversées qui rassemblent les différents termes d'indexation choisis pour représenter les contenus des documents et les liens vers ces documents (Hubert et al., 2009). En complément, à chaque couple (terme d'indexation, document) est associé un poids qui représente l'importance du terme dans un document. Lorsqu'une requête est soumise au système, les termes qu'elle contient sont mis en correspondance avec les termes

[1] http://www.dodiesis.com

d'indexation extraits des documents pour en déduire les documents à restituer à l'utilisateur. La phase d'indexation est donc une phase primordiale dans le processus de recherche.

Dans cette section, nous présentons dans un premier temps les méthodes d'indexation automatique des documents, puis dans un deuxième temps, les différents modèles de recherche.

2.2.1. Indexation automatique de documents

L'indexation des documents consiste à construire les descripteurs qui représentent chacun des documents, dans le but de pouvoir les rechercher ultérieurement. Chacune de ces représentations sera ensuite comparée à celle de la requête pour déterminer les documents potentiellement pertinents. L'indexation est une étape primordiale puisque la qualité de la restitution des documents dépendra de la qualité de l'indexation.

L'indexation peut-être réalisée manuellement (indexation analytique) et s'appuyer sur un thésaurus ou automatiquement. Nous détaillons dans les sections suivantes les principes des méthodes automatiques d'indexation par sac de mots et l'indexation par métadonnées.

2.2.1.1. Indexation par sac de mots.

L'indexation automatique des textes comprend deux étapes : la recherche des termes caractérisant le contenu et l'évaluation du pouvoir de caractérisation de ces termes. Différents problèmes sont à résoudre (Mothe, 1994) :

- définir l'élément qui sera choisi comme unité d'indexation (radical, mot simple, groupe de mots),
- choisir les termes représentatifs du document et ceux qui ne le sont pas, en fonction du contenu du document (termes d'indexation),
- évaluer le pouvoir de caractérisation de ces termes : certains termes sont plus importants que d'autres dans la caractérisation du contenu.

Au cours de ces différentes étapes, différents types de traitements sont appliqués qui peuvent être de type linguistique ou de type statistique.

L'indexation par sac de mots utilise le vocabulaire issu des documents (extrait par analyse des documents). A partir des mots utilisés dans le texte, l'indexation va s'appuyer sur différents éléments et techniques qui sont développés dans les paragraphes suivants :

- l'utilisation d'un anti-dictionnaire en fonction de la langue du texte,
- la troncature ou la radicalisation (*stemming* en anglais) pour déterminer une unité d'indexation unique pour différentes formes d'un mot (variantes morphologiques).
- la pondération pour représenter le fait que tous les termes ne représentent pas avec la même force le contenu du document.

a) Anti-dictionnaire

Pour éviter de retenir comme termes d'indexation des termes qui ne correspondent pas à la sémantique du texte c'est-à-dire qui ne définissent pas les thèmes traités par le document, un anti-dictionnaire (*stop list* en anglais) peut être utilisé. Un anti-dictionnaire contient les mots vides (articles, pronoms, prépositions, locutions, adjectifs démonstratifs, relatifs et possessifs, verbes auxiliaires, mots outils,...) et les mots athématiques c'est-à-dire qui se retrouvent dans n'importe quel texte indépendamment de son contenu. Ces termes ne sont pas intéressants

pour l'indexation dans la mesure où leur présence dans la plupart des textes ne permet pas de discriminer, de partitionner, pour une requête, les textes pertinents des textes non pertinents.

b) Troncature et radicalisation

Il peut être intéressant de représenter les différentes formes d'un mot (variantes morphologiques) par une même unité d'indexation, qui correspond à un radical. Différentes techniques sont utilisées pour cela, issues de considérations statistiques et de considérations linguistiques.

Pour la langue anglaise l'algorithme de PORTER (Porter, 1980) basé sur ce principe est un des plus utilisés. Différentes versions en différentes langues sont disponibles. Plus de détails se trouvent sur le site de Martin Porter (http://www.tartarus.org/~martin) en particulier dans le module « Algorithme de radicalisation de PORTER. D'autres langues sont traitées dans le projet à l'initiative de Martin PORTER, (http://www.snowball.sourceforge.net) : français, espagnol, portugais, italien, allemand, suédois, norvégien, danois et russe.

c) Pondération

La pondération des termes d'un document vise à considérer l'importance donnée aux termes caractéristiques de chaque document. La plupart des moteurs d'indexation utilisent des éléments statistiques pour pondérer l'importance d'un terme en fonction du nombre d'occurrences du terme dans le document (fréquence relative) et de la fréquence d'apparition de ce terme dans le corpus (fréquence absolue). Le poids de chaque terme pour un document représente le pouvoir de discrimination du terme pour ce document. Les termes de fréquence absolue faible permettront de distinguer, pour une requête, les documents pertinents des documents non pertinents. Dans le cas contraire, les réponses obtenues en utilisant des termes de fréquence absolue élevée comporteront beaucoup de documents et le risque que beaucoup soient non pertinents est important.

Ainsi, pour effectuer une recherche efficace, il est important de pouvoir prendre en compte le rôle de caractérisation des différents termes d'indexation à la fois dans un document et dans la collection. Des études statistiques sur la fréquence d'apparition des termes sont donc nécessaires afin d'associer à chacun des termes d'indexation une pondération représentative de la fréquence d'apparition d'un terme à la fois dans un document et dans la collection de documents.

Statistiquement, il est considéré que l'importance d'un terme est proportionnelle à sa fréquence relative et inversement proportionnelle à sa fréquence absolue (Sparck Jones, 1972).

La fréquence absolue inverse d'un terme, qui mesure l'importance d'un terme dans l'ensemble de la collection, peut s'écrire (Sparck Jones, 1972) :

$$idf_j = Log\left(\frac{N}{f_j}\right)+1 \qquad \textbf{(1)}$$

où N est le nombre total de documents dans la base et f_j est le nombre de documents qui contiennent le terme t_j. N/fj (1..N) car on a toujours $fj \neq 0$.

Cette formule permet de mettre en valeur les termes qui apparaissent dans peu de documents. Le poids d'un terme dans un document (Sparck Jones, 1972) est alors :

$$Poids_i(j) = tf_{ij} \cdot idf_j \quad \textbf{(2)}$$

où tf_{ij} est la fréquence d'apparition du terme t_j dans le document i.

La taille des documents, ainsi que la partie du document d'où sont extraits les termes (gros titre, titre, corps du document, etc) sont considérés dans des formules afin de prendre en compte l'importance des termes utilisés par les auteurs dans différentes parties du document.

Par ailleurs, un coefficient de pondération plus important peut également être associé à des termes d'une certaine catégorie grammaticale (les noms plus que les verbes), ou aux termes issus d'une liste ou respectant une syntaxe particulière. Enfin, l'indexation d'un document peut se faire aussi bien sur l'ensemble du document que sur une partie bien définie seulement.

2.2.1.2. Indexation par métadonnées

Une des méthodes proposées pour améliorer la qualité de la RI consiste à accompagner les documents d'un ensemble de métadonnées. Elles sont destinées à caractériser les documents et correspondent à l'ensemble des informations techniques et descriptives des documents comme les noms des auteurs, la date de publication, les mots clés, l'éditeur, les droits etc…

Selon le type des documents, des standards ont été développés afin de pouvoir partager ces métadonnées. Le Dublin Core Metadata Initiative (DCMI ou Dublin Core) (http://www.dublincore.org/) est un schéma de métadonnées générique qui permet de décrire des ressources numériques ou physiques et d'établir des relations avec d'autres ressources. Il comprend officiellement 15 éléments de description formels (titre, créateur, éditeur), intellectuels (sujet, description, langue, …) et relatifs à la propriété intellectuelle. Les métadonnées sont mises en ligne, au sein même des documents du corpus ou dans des fichiers spéciaux eux-mêmes accessibles dans le corpus.

La norme LOM (Learning Object Metadata), à son tour, est un schéma de description de ressources d'enseignement et d'apprentissage. LOM peut être utilisé pour décrire des ressources tant numériques que non numériques. LOM est détaillé dans la section 2.6.3.1. Les métadonnées, par leur définition même, apportent des informations de nature sémantique sur les documents qu'elles décrivent. Leur représentation à travers d'une ontologie peut être une solution intéressante pour spécifier et interpréter la sémantique des métadonnées.

2.2.2. Modèles de recherche

Pour évaluer l'appariement entre une requête et les documents, plusieurs modèles de recherche peuvent être utilisés par les SRI. Dans un premier temps, nous présentons les modèles booléen, vectoriel et probabiliste. Puis, nous examinons les modèles de langages (LM, pour Language Model) et le modèle Indexation Sémantique Latente (LSI, pour Latent Semantic Indexing).

2.2.2.1. Modèles booléen, vectoriel et probabiliste

Le modèle de recherche ***Booléen*** est le premier et le plus simple des modèles. Il est fondé sur la théorie des ensembles et l'algèbre de Boole. Le principe est simple : chaque terme de la requête est soit présent ou absent dans le document, d'où les poids binaires des termes qui sont soit 0 ou 1. Ainsi, un document est soit pertinent soit non pertinent par rapport à une requête. Il est possible d'exprimer la requête à l'aide des opérateurs logiques tels que *And*, *Or*,

Not, etc (Van Rijsbergen, 1979). Dans ce cas, un document est pertinent si et seulement si son contenu respecte bien la formulation logique demandée par l'utilisateur.

L'avantage de ce modèle est qu'il est transparent et peut être compris par l'utilisateur. Non seulement il n'a pas de paramètre caché, mais également la raison de sélection d'un document est claire, c'est-à-dire lorsque ce dernier correspond à la formule logique exprimée par l'utilisateur. De ce fait, ce modèle est bien adapté aux spécialistes.

Cependant, il présente des inconvénients tels que :

- la difficulté d'expression des requêtes longues sous forme booléenne,
- la non efficacité de critère binaire (0 ou 1) par rapport à la pondération des termes qui améliore les résultats,
- l'impossibilité de classement des documents car tous les documents retournés sont tous pertinents de la même façon.

Toutefois, lorsque l'utilisateur n'a qu'une vague idée de son besoin en information, il est souhaitable qu'il puisse avoir accès à des documents ne répondant que partiellement à sa requête. Le ***modèle booléen étendu*** a été proposé dans (Salton et al., 1983) afin de permettre l'utilisation des opérateurs logiques tout en proposant une pertinence graduée. Pour ce faire, il introduit le poids des termes de la requête dans le calcul de similarité. De façon similaire, le ***modèle booléen flou*** permet de représenter une pertinence partielle dans l'appariement requête-document (Baranyi et al., 1998).

Le modèle ***Vectoriel*** a été proposé par (Salton et al., 1971). Comme son nom l'indique, dans ce modèle, les documents et les requêtes sont représentés par des vecteurs. Les coordonnées des vecteurs sont exprimées dans un espace euclidien à N dimensions où N représente le nombre de termes d'indexation utilisés dans l'ensemble du corpus. Chacune des coordonnées correspond au poids du terme associé. Ainsi, les documents sont représentés par une matrice de taille N×M avec M le nombre de documents et N le nombre total de termes d'indexation dans tous les documents.

La pertinence d'un document correspond avec le degré de similarité entre le vecteur de la requête et celui du document. Le principe de la mesure de similarité entre un document et une requête est basé sur le fait que plus les deux représentations (document, requête) contiennent les mêmes informations, plus elles sont supposées représenter la même information.

Ce modèle de recherche présente plusieurs avantages tels que:

- le langage de requête est plus simple car une liste de termes,
- les performances sont meilleures grâce à la pondération des termes,
- la restitution de documents à pertinence partielle est possible,
- la fonction d'appariement permet de trier les documents résultats.

Cependant, quelques inconvénients sont constatés sur ce modèle recherche:

- le modèle ne considère pas les éventuels liens qui peuvent exister entre les termes,
- le langage de requête est moins expressif,
- l'utilisateur voit moins pourquoi un document lui est renvoyé.

Le modèle de recherche ***Probabiliste*** est basé sur l'estimation de la probabilité de pertinence d'un document par rapport à une requête (Robertson, 1977). Le modèle probabiliste présente

des résultats comparables avec ceux du modèle vectoriel (Croft et al., 1992). Un inconvénient de ce modèle est aussi l'indépendance des termes.

2.2.2.2. Les modèles LM, LSI

(Ponte et Croft, 1998) ont introduit le modèle de langage (modèle basé sur l'analyse des n-grammes) qui calcule la probabilité que des séquences de mots apparaissent dans un document donné. En d'autres termes, le modèle de langage mesure la probabilité de générer la requête à partir du modèle de langage du document.

Ainsi, dans un modèle n-gramme, la probabilité $P(w_1,\ldots,w_m)$ de retrouver la séquence $w_1,\ldots,w_m$ dans le document est calculée par (Ponte et Croft, 1998) :

$$P(w_1,...,w_m) = \prod_{i=1}^{m} P(w_i \mid w_1,...,w_{i-1}) \approx \prod_{i=1}^{m} P(w_i \mid w_{i-(n-1)},...,w_{i-1}) \qquad \textbf{(3)}$$

La probabilité conditionnelle peut être calculée avec le comptage des fréquences des n-grammes.

$$P(w_i \mid w_{i-(n-1)},..., w_{i-1}) = \frac{count(w_{i-(n-1)},..., w_i)}{count(w_{i-(n-1)},..., w_{i-1})} \qquad \textbf{(4)}$$

L'indexation sémantique latente LSI (Deerwester *et al.*, 1990) (Dumais, 95), (Foltz, 90), (Furnas et al, 88) vise à déterminer le thème véhiculé dans des documents par l'analyse globale du document en ne s'appuyant pas uniquement sur les termes. Toutefois, LSI s'intéresse aussi aux mots, phrases qui sont sémantiquement proches des termes recherchés. Ainsi, LSI permet de trouver des documents pertinents même s'ils ne contiennent aucun terme de la requête. Ce modèle utilise une matrice qui contient les termes sur les lignes et les documents sur les colonnes mais qui sont représentés dans un espace de dimension réduite issu de l'espace initial des termes d'indexation (Deerwester et al., 90), (Berry et al., 95). Cette réduction de dimension se fait par regroupement des termes ayant des caractéristiques communes dans leur apparition dans les documents via la décomposition aux valeurs singulières (Lebart et al., 1997).

Par rapport au modèle vectoriel, LSI réduit la dimension de l'espace de représentation aux vecteurs de représentation de l'information sémantique tout en minimisant l'effet de variation d'utilisation des termes. Ce modèle donne un meilleur résultat que les modèles statistiques (Dumais, 1995).

2.2.3. Conclusion

Le domaine de la RI ne cesse d'évoluer. Différents modèles de recherche utilisant des techniques diverses ont été utilisés dans la littérature, mais le point commun de tous ces modèles est l'indexation des documents basée sur les termes présents dans les documents. Cependant, le modèle LSI tend à prendre en compte la notion de sémantique des termes dans son modèle en regroupant ceux qui ont des propriétés communes. En même temps, l'exigence

des utilisateurs en matière de qualité de réponse des SRI ne cesse non plus d'évoluer. Actuellement, la notion de RI sémantique dévient un sujet incontournable car elle doit permettre d'obtenir un meilleur résultat par rapport aux précédents modèles de recherche dans la mesure où elle s'intéresse d'avantage à la compréhension du message véhiculé dans les documents et dans la requête.

Nous présentons dans la section suivante les différentes techniques utilisées actuellement pour la mise en œuvre de la RI sémantique.

2.3. Recherche d'information sémantique

La RI sémantique a pour objectif de mieux répondre aux besoins en information en prenant en compte les sens des mots aussi bien du côté de la requête utilisateur que celui des documents des corpus. Les sens des mots sont définis dans une base de connaissance qui peut être une ***Ontologie.*** .

2.3.1. Ontologie, normes et langages

Une ontologie regroupe les concepts qui représentent l'ensemble des connaissances d'un domaine en une spécification explicite et formelle (Studer, 1998). Elle montre les relations ainsi que les règles d'associations qui existent entre ces concepts. Elle permet ainsi, d'une part à l'ordinateur la production de nouvelles connaissances par le biais d'inférences, et d'autre part à l'homme et à l'ordinateur d'accorder des sens communs aux termes utilisés dans un domaine d'activité afin de lever toute ambiguïté pendant les traitements (Hernandez et al., 2006).

La description d'une ontologie repose sur différentes normes qui sont présentées dans les sections suivantes.

2.3.1.1. RDF

RDF (Resource Description Framework) (W3C, 2004) : cette norme accroît les possibilités d'exploitation des métadonnées. Créée par le W3C (World Wide Web Consortium), RDF facilite l'insertion et le traitement des métadonnées. Les métadonnées sont des données à propos des données (par exemple, un catalogue de bibliothèque est une compilation de métadonnées, puisqu'il décrit les publications). Dans le contexte du traitement des documents sur la toile, les métadonnées sont des données décrivant les ressources disponibles. RDF peut être utilisé pour établir des ontologies.

RDF est basé sur un modèle pour représenter des propriétés et des valeurs de propriétés données. Un document RDF est un ensemble de triplets de la forme <sujet, prédicat, objet> qui signifie que le sujet ***a*** a comme valeur pour le prédicat (propriété) b l'objet c.

Les éléments de ces triplets (a,b,c) peuvent être des URIs (Universal Resource Identifiers) (Berners Lee et al. 1999), des littéraux ou des variables.

2.3.1.2. OWL

OWL (Web Ontology Language) (W3C, 2004) est un langage de description d'ontologies conçu pour la publication et le partage d'ontologies sur le web sémantique.

Le langage d'ontologie Web OWL est conçu pour des applications qui doivent traiter le contenu des informations plutôt que de simplement les présenter aux humains. Le langage OWL offre aux machines de plus grandes capacités d'interprétation du contenu Web que

celles permises par XML, RDF et le schéma RDF (RDF-S), grâce à un vocabulaire supplémentaire et une sémantique formelle.

Ce langage y ajoute plus de vocabulaire pour décrire les propriétés et les classes entre autres, les relations entre les classes, cardinalité, égalité, typage de propriétés plus riche, caractéristiques des propriétés et les hiérarchies des propriétés et des classes.

OWL possède des sous-langages de plus en plus expressifs OWL Lite, OWL DL et OWL Full.

- OWL Lite : Il a été conçu pour être utilisé dans des situations qui n'ont besoin que des hiérarchies de classification et des caractéristiques de contraintes simples. Pour les contraintes de cardinalité, seules les valeurs 0 et 1 sont permises.
- OWL DL : Il est plus expressif qu'OWL Lite et est basé sur la description logique. Il est destiné aux utilisateurs qui demandent un maximum d'expressivité tout en maintenant la complétude (garantie de calculer toutes les conclusions) et la décidabilité (tous les calculs doivent finir en un temps fini). OWL DL contient tous les constructeurs du langage OWL mais sont utilisables avec des restrictions.
- OWL Full : Il est le plus expressif des sous langages d'OWL. Il est destiné aux utilisateurs qui demandent un maximum d'expressivité avec la liberté syntaxique de RDF sans aucune garantie de calcul. Par exemple, une classe peut être traitée comme une collection d'individus et en même temps peut être vue comme un seul individu. OWL Full permet aussi à une ontologie d'augmenter le sens du vocabulaire prédéfini (RDF et OWL).

Ainsi, les fonctionnalités de OWL Lite sont inclues dans celles de OWL DL. De même, OWL Full inclut toutes les fonctionnalités de OWL DL.

En résumé, OWL Lite $\subset$ OWL DL $\subset$ OWL Full.

Suite au développement du web sémantique et de ses technologies (Berners-Lee, 2001), différents travaux s'intéressent à son application en RI. L'hypothèse sous-jacente est que l'utilisation des concepts d'ontologie comme vocabulaire de référence d'index servirait à bien préciser les sens accordés aux termes d'un document et permettrait de lever les ambiguïtés des sens des termes utilisés et de mieux représenter les connaissances issues des documents. Nous présentons ci-après l'indexation de documents à base d'ontologies.

2.3.2. Indexation à base d'ontologies

Ce type d'indexation utilise des ontologies pour représenter les connaissances d'un domaine et ainsi faciliter l'acquisition et la compréhension des informations dans les documents ; d'où l'appellation d'indexation sémantique. L'indexation sémantique se base sur l'hypothèse que le sens des informations textuelles se comprend à partir des relations conceptuelles existant dans le domaine dont parle le contenu du texte plutôt qu'à partir des relations linguistiques et trouvées dans les documents (Haav et Lubi, 2001). Ainsi, l'indexation sémantique nécessite l'utilisation de ressources extérieures représentant explicitement l'information correspondant aux concepts traités dans les textes. Dans la littérature, l'indexation sémantique peut se faire soit suivant l'approche issue de la RI, soit suivant celle du Web Sémantique (Hernandez et al., 2008).

L'approche issue de la RI consiste à choisir l'ensemble des concepts et instances de l'ontologie comme langage de représentation des documents. Cette utilisation d'ontologies en tant que hiérarchies de concepts est le prolongement de l'utilisation dans le cadre de la RI des ressources terminologiques (Haav et Lubi, 2001). En effet, une ontologie peut fournir le vocabulaire et sa sémantique, ainsi que la structure des métadonnées associées aux ressources annotées. Ainsi, les descripteurs sont choisis au sein d'un vocabulaire contrôlé que constitue l'ontologie plutôt que directement dans les documents. De ce fait, les documents sont alors indexés par des concepts qui reflètent leur sens plutôt que par des mots bien souvent ambigus (Aussenac et Mothe, 2004). Nous voyons ici l'utilité de se servir d'une ontologie lors de l'indexation qui reflète bien le domaine de connaissance abordé dans le corpus. Dans ce type d'indexation sémantique, les concepts ou instances de concepts sont pondérés pour chaque document en fonction de la structure conceptuelle d'où ils sont issus, après les avoir identifié dans les documents (Haav et Lubi, 2001).

L'approche issue du domaine du Web sémantique utilise également l'indexation sémantique pour arriver à ses objectifs qui sont d'ajouter au contenu du Web une structure formelle et de la sémantique dans le but de permettre une meilleure gestion et un meilleur accès aux informations. (Berners-Lee et al. 1999) considèrent que les ressources participant au Web sémantique seront toutes reliées entre elles par des relations sémantiques. En d'autres termes, les documents utilisés sur le Web sémantique seront représentés par des ontologies qui décrivent les connaissances du domaine traités dans les documents. Cette démarche se fait par d'une part l'annotation des contenus des documents à l'aide des ressources conceptuelles et d'autre part l'extraction des concepts et instances. L'annotation de documents a pour objectif d'apporter des informations supplémentaires décrivant chaque document. Ces informations supplémentaires peuvent être des renseignements relatifs au média (date de création, taille, format d'encodage…), des métadonnées présentes dans les documents (auteurs, date de production,…), des index (les descripteurs du contenu du document), l'identifiant du document par le système et une vue sur le contenu (résumé ou extraits) (Euzenat, 2002).

Du point de vue sémantique, l'indexation avec une ontologie permet d'exprimer les relations entre des expressions du document à l'aide de celles des concepts auxquels les expressions sont associées dans l'ontologie. En termes d'indexation sémantique, des concepts de l'ontologie sont associés à chaque document selon les sémantiques qui y sont véhiculées.

Différents travaux ont montré l'intérêt d'utiliser une indexation sémantique à base d'ontologie. Dans le domaine de l'apprentissage en ligne, (Chang et al., 2007) proposent une indexation basée à la fois sur une ontologie du domaine de l'apprentissage et sur une ontologie dérivée de LOM (Learning Object Metadata), qui représente les métadonnées décrivant les ressources pédagogiques. Dans le cadre des recherches d'objets pédagogiques relatifs aux mathématiques en secondaire, les résultats montrent une meilleure efficacité en termes de rappel et de précision par rapport aux mêmes recherches basées sur des mots-clés. De même, (Hernandez et al., 2008) utilise les termes d'une ontologie de domaine, associée à une ontologie de tâche et de scenario d'apprentissage comme valeurs des métadonnées de LOM. Afin d'accéder aux instances d'ontologie d'une part et aux index associés aux documents d'autre part, (Hernandez et al., 2007) proposent de les stocker dans une base de données relationnelles. Par ailleurs, (Song et al., 2005) propose un modèle de RI basé sur des ontologies de domaine, définies avec OWL lite. Les différentes ontologies de domaines, exprimées en OWL, sont intégrées pour former une ontologie unique. Les termes définis dans l'ontologie sont alors utilisés d'une part comme métadonnées pour annoter les contenus du web et d'autre part comme termes d'indexation de la collection.

Les documents une fois annotés avec des concepts d'ontologie peuvent être recherchés sémantiquement moyennant des fonctions de similarité sémantique qui évaluent la similitude

entre les concepts des documents avec ceux de la requête utilisateur. Nous présentons dans la section 2.3.3 les différentes mesures de similarité conceptuelle qui estiment la ressemblance entre deux concepts de l'ontologie. Puis dans la section 2.3.4 nous présentons les différentes techniques de pondération de ces similarités conceptuelles afin d'accorder plus d'importance à certains concepts par rapport aux autres. Ensuite, nous abordons l'évaluation de la requête avant de traiter les fonctions de similarité sémantique entre requête et document.

2.3.3. Mesures de similarité conceptuelle

La mesure de la proximité conceptuelle sur des réseaux sémantiques remonte aux travaux de (Quillian, 1968) et à ceux de (Collins et Loftus, 1975) sur la mémoire sémantique humaine. Pour estimer un degré d'adéquation entre une requête et un document, il est nécessaire de disposer d'une mesure de proximité sémantique qui permet de calculer le degré de ressemblance essentielle de deux concepts par une valeur numérique, en vue de classer les documents potentiellement pertinents par ordre décroissant de pertinence.

Deux grandes familles d'approches peuvent être identifiées pour le calcul de telles distances : celles qui incluent des informations externes à la hiérarchie, par exemple, des statistiques sur l'utilisation des types de concepts (Resnik, 1995) (Jiang et Conrath, 1997), et les approches reposant uniquement sur la structure hiérarchique de l'ontologie (Rada et al., 1989) (Wu et Palmer, 1994).

La mesure Edge Counting (Rada et al., 1989) présente une mesure utilisant une métrique, dist(c1, c2), qui évalue le nombre d'arcs minimum à parcourir pour aller d'un concept c1 à un concept c2. Cette mesure indique la distance sémantique entre ces deux concepts.

La similarité sémantique entre deux concepts correspond à l'inverse de la distance entre deux concepts. Plus deux concepts sont distants, moins ils sont similaires.

$$Sim_{Rada}(c_1, c_2) = \frac{1}{1 + dist_{edge}(c_1, c_2)} \qquad \textbf{(5)}$$

Où la distance $dist_{edge}(c_1, c_2)$ est la longueur du plus court chemin entre deux concepts c_1 et c_2.

À partir de la mesure de distance précédente, (Leacock, 1998) a proposé une formule pour calculer la similarité. Elle est issue de la proposition de (Resnik, 1998).

$$Sim_{edge}(c_1, c_2) = -\log\left(\frac{dist_{edge}(c_1, c_2)}{2 * Max}\right) \qquad \textbf{(6)}$$

Où Max désigne la profondeur maximale de la taxonomie.

D'autres mesures utilisent la notion de plus petit généralisant commun, c'est-à-dire le généralisant commun à c1 et c2 le plus éloigné de la racine. Ainsi la mesure de (Wu et Palmer, 1994) est :

$$Sim_{wp}(c_1, c_2) = \frac{2 * depth(c)}{depth(c_1) + depth(c_2)} \qquad \textbf{(7)}$$

où depth(c_i) correspond au niveau de profondeur du concept c_i dans la hiérarchie et c est le concept subsumant c1 et c2.

Tout à fait différemment, des approches basées sur les nœuds, cherchent le contenu informatif des nœuds. Deux versions existent :

- La première utilise un corpus d'apprentissage et mesure la probabilité de trouver un concept ou un de ses descendants dans ce corpus. Soit c un concept, et p(c) la probabilité de le trouver lui ou un de ses descendants dans le corpus. Le contenu informatif associé à c est alors défini par :

$$IC(c) = -\log(p(c)) \qquad \textbf{(8)}$$

$$\text{Avec } p(c) = \frac{freq(c)}{N} \quad \textbf{(9)} \quad \text{et } Freq(c) = \sum_{n \in Word(c)} Count(n) \quad \textbf{(10)}$$

Où word(c) est l'ensemble des termes ou labels représentant le concept c et les concepts subsumés par c, count(n) est le nombre d'occurrences du terme n dans le corpus et N le nombre total d'occurrences des labels de concepts retrouvés dans le corpus.

La proximité entre les concepts c_1 et c_2 nécessite de trouver l'ensemble des concepts qui les subsument tous les deux. Soit $S(c_1,c_2)$ cet ensemble. Selon (Resnik,1995), la mesure de similarité est :

$$Sim_{Resnik}(c_{1,}c_2) = Max(IC(c)), c \in S(c_1, c_2) \qquad \textbf{(11)}$$

Où IC(c) est le contenu informatif associé au concept *c* défini en (8).

- La seconde version refuse l'utilisation d'un corpus et essaie de calculer le contenu informatif des nœuds à partir de WordNet (Felbaum,1998) uniquement.

L'hypothèse de (Seco et al.,2004) est que, plus un concept possède de descendants, moins il est informatif. Ils utilisent donc les hyponymes des concepts pour calculer le contenu informatif de ceux-ci, comme suit :

$$IC_{wn}(c) = \frac{\log\left(\frac{hypo(c)+1}{\max_{wn}}\right)}{\log\left(\frac{1}{\max_{wn}}\right)} = 1 - \frac{\log(hypo(c))+1}{\log(\max_{wn})} \qquad \textbf{(12)}$$

avec hypo(c) qui indique le nombre d'hyponymes dont dispose le concept c, et max_{wn} qui indique le nombre de concepts de la taxonomie. Les différentes mesures de similarité sémantique utilisant le contenu informationnel de (Resnik,1995) peuvent donc être redéfinies en utilisant celui de (Seco et al., 2004).

Les deux grandes approches définies précédemment peuvent être combinées. Souvent, il s'agit de réutiliser le contenu informatif et le plus petit ancêtre commun (c), comme avec Lin (1998) :

$$Sim_{Lin}(c_1, c_2) = \frac{2 * \log P(c)}{\log P(c_1) + \log P(c_2)} \qquad \textbf{(13)}$$

ou avec (Jiang et Conrath, 1997) :

$$Sim_{Jiang-Conratg}(c_1, c_2) = IC(c_1) + IC(c_2) - 2 * IC(c) \qquad \textbf{(14)}$$

Les similarités conceptuelles peuvent être pondérées suivant les importances relatives entre les différents types de concepts. Nous présentons ci-après les techniques de pondération des similarités entre concepts.

2.3.4. Pondération des concepts et instances

Le calcul du poids d'un concept ou d'une instance dans la représentation d'un granule peut être fait suivant plusieurs approches : statistiques ou conceptuelles.

- Pondération statistique : L'approche proposée dans (Vallet et al., 2005) a pour but de calculer le poids des instances. Elle est inspirée de la méthode tf.idf. Le poids $w_{i,j}$ d'une instance I_i dans un document D_j est calculé ainsi :

$$w_{i,j} = \frac{freq_{i,j}}{\max_k freq_{k,j}} * \log \frac{N}{n_i} \qquad \textbf{(15)}$$

où $freq_{i,j}$ représente le nombre d'occurrences de I_i dans D_j, $\max_k freq_{k,j}$ est la fréquence de l'instance dans D_j, n_i est le nombre de documents annotés avec I_i et N est le nombre total de documents dans la collection.

Le nombre d'occurrences d'une instance a été défini comme le nombre de fois où un label de l'instance apparaît dans le texte, si ce document est annoté avec l'instance, ou 0 s'il ne l'est pas. Cependant, les résultats obtenus n'ont pas été satisfaisants car un grand nombre d'instances n'était pas reconnu à cause de lacunes à l'étape d'extraction des labels (non prise en compte des pronoms et périphrases notamment).

Une approche similaire est présentée pour la pondération de concepts dans (Baziz et al., 2005). L'inconvénient de ces approches est qu'elles ne considèrent que les occurrences des concepts ou instances dans les documents et ne considèrent pas l'organisation conceptuelle dont ils sont issus. Une partie de la sémantique contenue dans les relations entre concepts est alors ignorée. D'autres approches visent à combiner la pondération des concepts et/ou instances à partir de leurs occurrences dans les documents et leur place dans la représentation conceptuelle.

- Pondération conceptuelles : Dans (Desmontils et Jaquin, 2002) une approche est présentée pour indexer un ensemble de sites Web à partir d'une ontologie. Le pouvoir représentatif d'un concept prend en compte la fréquence d'apparition des termes désignant le concept dans les sites mais également ses relations avec les autres concepts du domaine.

Plus un concept possède de relations avec les autres concepts présents dans la page, plus il est représentatif de la page. Le pouvoir se calcule de la façon suivante : les termes d'une page Web sont tout d'abord extraits après analyse syntaxique avec tree tagger[2] à partir de patrons (nom, nom+nom, nom+adjectif). Un premier poids, appelé poids de fréquence est calculé pour chaque terme en fonction de sa fréquence d'apparition et des balises HTML qui l'encadrent.

Les coefficients correspondant à chaque balise sont attribués expérimentalement : par exemple, si un terme est encadré par la balise titre, le coefficient est 10, s'il est mis en gras, le coefficient est 2. En supposant qu'un terme T_i apparaît p fois dans une page contenant n termes, Mi étant le coefficient relatif à la balise encadrant l'occurrence j du terme T_i, le poids de fréquence P_freq de T_i est calculé ainsi :

$$P\_freq(T_i) = \frac{P(T_i)}{\max_{k=1..n}(P(T_k))} \quad \text{et} \quad \mathrm{P}(\mathrm{T_i}) = \sum_{i=1}^{p}(Mi, j) \qquad \textbf{(17)}$$

Ensuite, à partir de WordNet3, l'ensemble des concepts relatifs à ces termes est généré sous forme de synset en prenant tous les sens définis. Un poids, appelé poids sémantique, est ensuite calculé en mesurant la similarité entre le concept donné et l'ensemble des autres concepts retrouvés.

Pour calculer le poids sémantique d'un concept dans une page, la somme des mesures de similarité du concept avec les autres concepts retrouvés de la page est calculée de la façon suivante :

$$P\_sem(synset_i(T_k)) = \sum_{j\in[1,k-1]\cup[k+1,m]} \sum_{l=1}^{k} sim(synset_i(T_k), synset_l(T_j)) \qquad \textbf{(18)}$$

où $synset_i(T_m)$ représente le sens i dans WordNet retrouvé pour le terme T_m. Enfin, le pouvoir représentatif Rep du concept ou synset correspondant aux termes T_k est calculé en fonction de son poids sémantique et de son poids de fréquence :

$$\mathrm{Rep}(\mathrm{synset}\ (\mathrm{T_k})) = \frac{\alpha * P\_freq(T_k) + \beta * P\_sem(symset(T_k))}{\alpha + \beta} \qquad \textbf{(19)}$$

α et β sont fixés empiriquement à 1 et 2.

Les concepts retenus pour indexer chaque page sont ensuite choisis à partir d'un seuil sur ce pouvoir et de la présence de ce concept dans l'ontologie choisie pour indexer le corpus.

[2] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

[3] WordNet est une base de données lexicale développée par l'université de Princeton. Elle est disponible à l'adresse http://wordnet.princeton.edu/

2.3.5. Évaluation de requête

Dans le cas où les documents sont représentés à partir d'ontologies contenant des règles d'inférences (ontologies lourdes), des moteurs d'inférences peuvent être intégrés au système afin d'interroger la base de connaissance constituée des ontologies et des documents. Racer (Haarslev et Möller, 2001) et FaCT DL (Horrocks, 1998) sont des exemples de moteurs d'inférence reposant sur la logique de description. Afin d'interroger ces moteurs, plusieurs langages d'interrogation ont été définis à partir du langage formalisant la connaissance (Karvounarakis et al., 2002) (McBride, 2001) (Miller et al., 2002) (Guha et al., 2003). Ces langages fournissent des mécanismes permettant d'exprimer des requêtes complexes. La requête est alors exécutée sur la connaissance représentée dans l'ontologie et les instances qui satisfont la requête sont restituées. Les requêtes sont soit générées à partir d'une requête en langage libre (Guha et al., 2003) (Rocha et al., 2004), soit à partir d'interfaces permettant de sélectionner les classes et propriétés de l'ontologie qui intéressent l'utilisateur (Kiryakov et al., 2004), (Maedche et al., 2003). L'avantage de ce type d'interrogation est que des mécanismes d'inférence sont mis en place à partir de la classification hiérarchique des concepts et des règles. GetData présenté dans (Guha et al., 2003) permet d'interroger de façon simple et efficace une ontologie considérée comme un graphe étiqueté représenté en RDF. Il permet d'accéder à des ressources vérifiant une propriété.

L'avantage de ce type d'interrogation est qu'elle permet de mettre en place des procédés de raisonnement à partir des éléments de la requête et des éléments retrouvés dans les documents. Cependant, ce type d'appariement consiste à rechercher exactement les éléments présents ou inférés de la requête dans les documents. Les documents ne sont pas restitués par ordre de pertinence mais parce qu'ils contiennent les éléments cibles.

L'originalité de l'approche présentée dans (Vallet et al., 2005) est que l'ensemble des documents correspondant aux instances restituées pour une requête formulée en RDQL (RDF Data Query Language)[4] est ensuite classé par calcul de pertinence vis à vis de la requête. L'appariement repose sur la représentation vectorielle de la requête et des documents à partir des instances de l'ontologie, reprenant ainsi les principes développés par (Salton et al., 1983).

(Castells et al., 2007) soumettent un modèle pour l'exploitation de bases de connaissance basées sur une ontologie de domaine pour la recherche de documents dans un dépôt de documents de grande taille. Le modèle est basé sur l'adaptation du système vectoriel, en incluant un algorithme d'annotation pondéré, un algorithme de classement et la combinaison avec la recherche basée sur les mots-clés afin de combler d'éventuels manques ou incomplétudes dans la base de connaissance.

2.3.6. Fonction de similarité sémantique entre requête et documents

Les ontologies peuvent servir à calculer la similarité entre la représentation de la requête et la représentation des documents dans le cas où les deux représentations sont faites à partir des concepts d'une même ontologie.

Cette approche est suivie dans (Andreasen, 2003). Les documents et requêtes sont représentés à partir du langage et de l'ontologie Ontologu. Cette ontologie contient un ensemble de concepts et de relations entre concepts, dont la relation de subsomption. Elle est considérée comme un graphe orienté. L'avantage du calcul de la similarité est de classer les documents

[4] http://www.w3.org/Submission/RDQL/

restitués par rapport à leur similarité à la requête, cette similarité reposant sur l'organisation des concepts dans l'ontologie. Le calcul de similarité s'appuie sur trois intuitions :

- la première intuition est que les documents liés au concept généralisant ou spécifiant le concept utilisé dans la requête peuvent intéresser l'utilisateur. Le calcul de la similarité prend donc en compte la distance séparant les deux concepts par la relation de subsomption. La similarité revient à prendre le nombre d'arcs séparant les deux concepts par le chemin le plus court à partir de la relation de subsomption.
- la deuxième intuition est que deux concepts ayant un concept les généralisant (ou subsumeur) commun sont plus similaires. Afin d'appliquer cette intuition, chaque concept est représenté par un ensemble flou à partir des concepts le généralisant. La similarité entre concepts est alors calculée à partir des éléments faisant partie de l'intersection entre les descriptions des concepts.
- la troisième intuition est que la similarité entre concepts doit prendre en compte les relations autres que les relations de subsomption. L'ensemble des concepts généralisant les deux concepts est alors considéré. Un sous-graphe de l'ontologie est construit à partir des concepts de cet ensemble pouvant être reliés dans l'ontologie par n'importe quel type de relation. La similarité est calculée par rapport aux nombres de nœuds ainsi connectés.

Cette approche est originale car elle calcule la similarité entre les concepts des documents et les concepts de la requête. La mesure de similarité proposée repose sur l'organisation des concepts dans l'ontologie. Cependant, aucune indication n'est donnée sur la combinaison des différents facteurs de la mesure. De plus, les auteurs ne considèrent pas le cas de figure suivant lequel plusieurs concepts sont retrouvés à la fois dans la requête et les documents ni comment les différentes similarités sont combinées. Aucune évaluation n'est proposée.

Une autre approche est présentée dans (Guarino, 1999). Contrairement à la précédente, le mécanisme d'appariement est entièrement décrit ; cependant le procédé est manuel. Le but du système OntoSeek (Guarino, 1999) est d'améliorer l'accès aux pages jaunes à partir d'un mécanisme reposant sur WordNet. Les documents et les requêtes sont représentés à partir de graphes conceptuels formés de nœuds et d'arcs dont les labels sont issus de WordNet. Une interface aide l'utilisateur dans la conception de ces graphes. Pour étiqueter les nœuds, l'utilisateur peut soit proposer des mots qui sont ensuite désambiguïsés à partir de WordNet, soit directement naviguer dans WordNet pour sélectionner les synsets qui l'intéressent.

De la même façon, les arcs sont étiquetés soit à partir d'une liste proposée par le système, soit à partir de termes proposés par l'utilisateur. Un procédé d'appariement entre le graphe de la requête et l'ensemble des graphes représentant les documents est ensuite mis en place. Le système recherche les graphes de documents qui subsument (ou qui spécifient) le graphe de la requête. Les résultats sont présentés à l'utilisateur à partir d'une interface présentant un rapport en HTML.

(Zhao et al., 2007) utilisent un algorithme de similarité fondé sur un arbre sémantique construit à partir d'une ontologie de domaine OWTS (ontology-based weighted semantic tree similarity Algorithm) en se basant sur la représentation sémantique des requêtes et des titres de documents. (Aufaure et al., 2007) recherchent les informations pertinentes en appliquant les techniques du modèle vectoriel et le cosinus pour calculer la similarité entre concepts, après avoir reformulé les requêtes. (Xiaomeng et Atle, 2006) proposent une méthode heuristique semi-automatique qui permet de faire correspondre deux ontologies. La mise en correspondance entre ontologies est focalisée sur la mise en correspondance des concepts et relations où la distance sémantique est enrichie à l'aide de la lemmatisation, la recherche des

fonctions grammaticale des mots et les types (nom, verbe, adjectif…) des mots moyennant Wordnet. (Gligorov et al., 2007) proposent une méthode de mise en correspondance approximative de concepts basée sur la mesure de similarité utilisée par Google. Cette mise en correspondance approximatif est surtout nécessaire quand les concepts sont vagues, flous ou mal définis. L'appariement d'ontologie est souvent la recherche d'équivalence A ≡ B entre deux concepts A et B de deux différentes hiérarchies. L'équivalence est traitée comme une subsomption mutuelle entre A et B. A≡B si A⊆B et B⊆ A. La représentation des concepts comme une conjonction de concepts implique que les concepts ont la forme $B=B_1 \cap \ldots..\cap B_k$. Ainsi A⊆B si et seulement si $A \subseteq B_i$, $\forall$ i=1 à k.

(Gligorov et al., 2007) définissent l'approximation de façon à ce que des insatisfactions aux sous problèmes $A \subseteq B_i$ soient autorisées tout en disant que le problème initial doit être satisfait.

Le degré d'approximation est donné par le ratio entre le nombre de sous problèmes non satisfiables et le nombre total des sous-problèmes avec pondération heuristique de chaque sous problème avec le NGD (Distance Normalisée Google)

$$\text{Avec } Ngd(x,y) = \frac{\max(\log f(x), \log f(y)) - \log f(x,y)}{\log M - \min(\log f(x), \log f(y))} \quad \textbf{(20)}$$

Où f(x) est le nombre de documents pertinents trouvés par Google pour la recherche du terme x,

f(y) est le nombre de documents pertinents trouvés par Google pour la recherche du terme y,

f(x,y)) est le nombre de documents pertinents trouvés par Google pour la recherche du n-uplet(x,y),

M est le nombre de pages web indexées par Google.

Ces fonctions de similarité sémantique peuvent restituer des documents potentiellement pertinents par rapport aux requêtes des utilisateurs. Cependant, les résultats peuvent ne pas satisfaire l'utilisateur. Ainsi, afin d'améliorer les résultats de la RI, une reformulation peut être appliquée à la requête initiale. Nous présentons dans la section suivante les principes et techniques de reformulation des requêtes dans la littérature.

2.3.7. Reformulation des requêtes

Il a été prouvé que la reformulation de requêtes a des effets positifs en RI (Harman, 1992). L'objectif de la reformulation est soit de limiter le silence (le silence fait référence aux documents pertinents mais qui ne sont pas retrouvés par le système), soit de réduire les risques de bruit (le bruit fait référence aux documents non pertinents retrouvés par le système). Dans le premier cas, la requête est étendue à partir de termes similaires à ceux de la requête initiale. Dans le second cas, la requête initiale est étendue ou modifiée à partir de termes qui ajoutent de l'information complémentaire à la représentation du besoin. Alternativement, des principes de désambigüisation peuvent être appliqués.

Il y a principalement deux approches permettant l'expansion de requêtes. La première consiste à utiliser des ressources, comme par exemple un dictionnaire (Moldovan, 1999) ou

bien WordNet (Voorhes, 1994), en étendant les requêtes à partir de nouveaux termes en relation avec les termes de la requête. La deuxième solution est la ré-injection de pertinence reposant sur l'analyse des termes contenus dans les documents jugés pertinents pour la requête initiale. Cette approche, ne faisant pas intervenir d'ontologies, ne fait pas partie de notre étude.

Un autre intérêt des ontologies est de permettre la désambiguïsation des termes de la requête. Dans (Guha, 2003) la désambiguïsation se fait selon trois approches. La première consiste à choisir le concept dont les labels apparaissent le plus souvent dans les documents. La seconde approche consiste à réaliser un profil utilisateur et à choisir le concept le plus proche de son profil. Finalement, la troisième prend en compte le contexte de la recherche et les documents recherchés par l'utilisateur jusque là.

(Köhler et al., 2006) améliorent la désambigüisation des sens des mots en utilisant la lemmatisation des mots au lieu d'utiliser la radicalisation. De plus, ils proposent une méthode pour améliorer le rappel sans altérer la précision par l'utilisation des sous-concepts et super-concepts dans les différentes relations en respectant une certaine limite sur la profondeur des relations de subsomption.

Dans $(Ka)^2$ (Benjamins et al., 1999), les pages Web sont annotées manuellement par des concepts d'une ontologie. Pour une requête donnée, tous les concepts liés aux termes de la requête sont inférés et ajoutés à la requête. Une interface a été développée pour assister l'utilisateur dans la formulation ou le raffinement de sa requête. Elle repose sur la visualisation de l'ontologie à partir de vues hyperboliques. Il est ainsi possible de naviguer dans l'ontologie et de centrer la visualisation sur la représentation des concepts qui intéressent l'utilisateur comme il a été fait dans WebBrain[5].

(Aufaure et al., 2007) adoptent l'adaptation du modèle vectoriel en substituant les termes de la requête par des concepts de l'ontologie et en classifiant les résultats d'une requête par service. Les requêtes des utilisateurs sont enrichies par analyse morphologique et sémantique en utilisant les concepts et les relations entre l'ontologie de domaine et WordNet. L'utilisateur peut aussi utiliser l'ontologie de domaine pour choisir les concepts à utiliser dans sa requête.

(Tomassen et al., 2006), à leur tour, proposent d'enrichir la requête de l'utilisateur par substitution des concepts de la requête par les vecteurs caractéristiques des concepts correspondants dans l'ontologie. A chaque concept de l'ontologie de domaine est donc associé un vecteur caractéristique décrivant la similarité sémantique du concept avec les termes et concepts auxquels le concept est en relation (synonyme, conjugaison, etc) par rapport aux contenus des documents d'un corpus.

Dans (Kim et al., 2007), la recherche d'objets se fait en deux phases. D'abord, la requête de l'utilisateur est reformulée en cherchant dans l'ontologie les concepts qui correspondent aux mots clés de la requête. Après le classement sémantique des concepts retrouvés, le système réalise la recherche d'objets contenant ces concepts.

2.3.8. Conclusion

Différentes mesures de similarité ont déjà été développées dans la littérature. Les unes s'orientent vers la similarité conceptuelle se basant sur la hiérarchie des concepts présents dans les ontologies, les autres se basent sur des informations statistiques sur des occurrences des termes représentant les concepts. Ces dernières ont l'inconvénient de ne pas bien refléter

[5] http://www.Webbrain.com/html/default_win.html

la sémantique des concepts car les similarités conceptuelles varient d'un corpus à l'autre sur une même ontologie voire sur les mêmes corpus et ontologie mais sur deux versions différentes des documents dans le temps. Pour les mesures de similarité conceptuelle se basant sur la hiérarchie des concepts, nous constatons la non-conformité des résultats des mesures avec la distance sémantique entre les concepts de l'ontologie car des concepts se trouvant dans deux branches taxonomiques différentes ont des similarités assez élevées alors que ces dernières devraient être proches de zéro. Un manque de fonctions de similarité qui calculent la similarité entre graphes de concepts est aussi constaté. En effet, plusieurs types de concept peuvent être utilisés pour annoter un document. Cependant, ces différents types de concept peuvent ne pas avoir les mêmes degrés d'importance pour une ontologie donnée. Les fonctions de similarité entre graphes de concepts doivent donc tenir compte de ces importances relatives entre types de concept. Ainsi, nous avons défini des mesures de similarité conceptuelle ainsi qu'une mesure de similarité de graphes de concepts que nous présenterons dans le chapitre 3 (sections 3.4.3.1 et 3.4.3.7).

2.4. Dynamique en recherche d'information

Dans cette section, nous présentons les travaux relatifs à l'indexation dynamique des documents consécutifs aux modifications apportées dans la collection après avoir présenté un état de lieu sur la dynamique en RI.

2.4.1. État des lieux

La mise à jour des documents de la collection, l'arrivée de nouveaux documents ou la suppression de documents d'un corpus indexé nécessitent l'actualisation de l'index afin de garder la cohérence entre les documents et les index. Cette mise à jour est primordiale pour que le SRI puisse répondre au mieux aux besoins d'un utilisateur.

Lorsque la collection de documents est figée, l'indexation est réalisée une fois pour toutes. Cependant, ce cas n'intervient que dans le cadre des campagnes d'évaluation des moteurs où il s'agit de confronter différents SRI sur les mêmes collections. Dans l'usage réel, le SRI doit être capable de faire face à des collections dynamiques dans lesquelles des documents sont modifiés, ajoutés et supprimés.

La toile correspond à ce cadre de collections hautement dynamiques. Ainsi, des travaux dans le domaine de la recherche et de la collecte incrémentale des pages Web visent à permettre aux collections du moteur de recherche d'être synchronisées avec le Web réel. Dans (Cho et Garcia-Molina, 2000) la collecte vise à télécharger toutes les pages web relatives à une URL de départ. Ensuite, la collecte incrémentale ne télécharge que les pages qui ont été modifiées à la source. Dans cette méthode, la maintenance de la collection ne nécessite pas de télécharger les pages qui n'ont subi aucune modification.

Cependant, la collection synchronisée, qui est l'image exacte des documents sources, ne peut pas être recherchée immédiatement car la reconstruction (à partir de rien) de l'index de mots-clés est moins fréquente que la mise à jour de la collection. Il n'est donc possible de rechercher les nouveaux documents collectés qu'après la prochaine reconstruction de l'index. Le décalage entre la mise à jour de la collection et celle de l'index s'explique par le fait que la reconstruction complète de l'index est très coûteuse en temps de traitement.

2.4.2. Démarches dans la littérature

Face à ce problème, (Lim et al., 2007) proposent une méthode de mise à jour incrémentale d'index inversé pour les documents qui ont changé sur le web. Cette méthode qui s'appelle *délimiteur-diff* (Landmark-diff en anglais) combine la technique de *délimiteur* (Landmark en anglais) (Salton et al, 1993) avec la méthode *diff* (Ukkonen, 1985). Ces méthodes considèrent une indexation de type « sac de mots », dans laquelle les termes issus des documents sont considérés comme indépendants. La technique de délimiteur consiste à subdiviser les documents en plusieurs blocs et à mémoriser les positions relatives des mots du document par rapport aux délimiteurs du bloc dans lequel les mots se trouvent. La méthode *diff* par contre mémorise la liste des modifications apportées dans un document pour obtenir sa nouvelle version. A chaque document est donc associé un annuaire de délimiteurs qui liste les différents délimiteurs et leurs positions absolues dans le document. La mise à jour d'un document entraine celle de l'annuaire des délimiteurs associé au document. Ce nouvel annuaire des délimiteurs associés à la *transcription des modifications* (edit transcript en anglais) permet de mettre à jour l'index inversé. La transcription des modifications correspond à la liste des modifications qui amènent vers la nouvelle version d'un document à partir de l'ancienne version.

Une entrée dans l'index inversé est composée de l'identifiant de chacun des mots (wordID), la liste des documents contenant un mot donné (docID), l'identifiant du délimiteur (landmarkID) auquel le mot est rattaché, ainsi que la position relative (offset) du mot par rapport au délimiteur. La méthode Délimiteur-diff présente une vitesse de mise à jour de l'index inversé (pour les documents qui ont changé) trois fois plus rapide que la méthode *Premier Index* (forward index en anglais) utilisée par Google (Page et Brin, 1998).

La technique *Merge-based* (Cutting et Pedersen, 1990), quant-à elle, consiste à minimiser le déplacement de la tête de disque pour la maintenance de l'index. La mise à jour des index sur disque se fait dès que l'espace mémoire alloué aux index commence à être saturé. L'inconvénient de la technique *Merge-based* est que la totalité de l'index inversé est lu et écrit sur le disque à chaque mise à jour, même si une petite partie seulement de l'index est affecté. Les accès au disque sont donc importants.

La stratégie *In-Place* (Tomasic et al., 1994) consiste non seulement à transformer les structures de données en d'autres structures plus petites mais aussi à écraser les anciennes versions de documents. Cette stratégie essaie de résoudre ce problème en laissant assez de place à la fin des index inversés. Dès lors, (Büttcher et Clarke, 2006) proposent une approche hybride qui combine les deux méthodes *In-place* et *Merge-based* respectivement pour une longue liste inversée et pour une courte liste inversée. Cette approche hybride donne une meilleure performance en termes de temps d'indexation que l'une ou l'autre de ces méthodes, tout en gardant la même performance au niveau de traitement de la requête.

La plupart des algorithmes d'actualisation d'index ne permet pas l'ajout de nouveaux documents pendant le processus de mise à jour de l'index. De plus, ce processus peut demander plusieurs heures pour les corpus de grande taille.

(Galambos, 2006) a développé un algorithme de mise à jour dynamique d'index. Cet algorithme est basé sur la mise à jour incrémentale d'index en utilisant une liste inversée de type *Citerne (Tanker).* Dans ce modèle, une *citerne* est un index composé de *Barils* (*barrels*) ; où un baril est un index réalisé sur un sous-ensemble de documents du corpus. L'actualisation d'index est réalisée à chaque arrivée de nouveaux documents et à chaque modification de documents. La modification d'un document est considérée comme une opération de suppression suivie d'un ajout d'un nouveau document.

En ce qui concerne Google (Page et Brin, 1998), il emploie plusieurs techniques (importance des pages ou *PageRank*, structure des liens, texte des liens, polices de caractères, position des mots dans les documents) pour améliorer la qualité de recherche. L'analyse des structures des liens à partir de la popularité des pages permet à Google d'évaluer la qualité des pages web (Page et Brin, 1998). Pour atteindre ses objectifs, les structures de données (cf. figure 1) utilisées par Google sont :

- Liste d'importance (Hit list) : liste des occurrences d'un mot d'un document particulier, comprenant les informations de la position, la police et la taille (visuelle) du mot,
- Baril (Barrel) : index partiellement trié par le docID, obtenu par la méthode Premier Index (Forward-Index). Chaque baril stocke des listes d'importance pour un ensemble d'identifiants de mots (wordID). Il y a deux types de barils : les barils courts qui contiennent les listes d'importance incluant le titre ou les ancres et les barils longs pour toutes les listes d'importance,
- Entrepôt (Repository) : contient le code HTML des pages Web. Chaque document est préfixé par son identifiant docID, sa longueur, et son URL,
- Index Doc : garde les informations concernant les documents telles qu'un pointeur vers le dépôt, le nombre de liens entrant provenant d'autres pages, le nombre de liens sortant qui pointent vers d'autres pages, l'état de chaque document ainsi que son URL,
- Lexique : table de hashage gérée en mémoire vive qui garantit l'appariement entre un mot et son identifiant (wordID),
- Ancre : stocke les destinations et les étiquettes des liens.

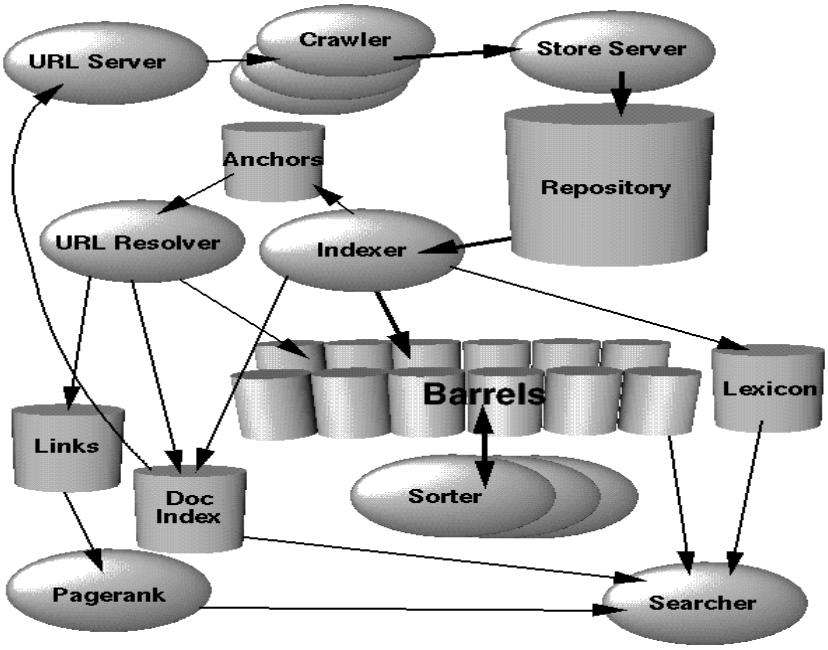

Figure 1 : les structures de données utilisées par Google (Page et Brin, 1998)

2.4.3. Conclusion

Les SRI sont généralement évalués au travers de leurs performances en termes de pertinences des documents retrouvés. Ils peuvent également être évalués par rapport au temps de réponse aux requêtes des utilisateurs ainsi que sur la disponibilité du système. Ainsi, plusieurs paramètres qui entrent en jeu dans la performance des SRI doivent être pris en compte. Parmi ces paramètres, nous pouvons citer : la taille du corpus (nombre de documents constituant la collection), la fréquence de mise à jour de la collection et le format des documents. En effet, la mise à jour de la collection peut demander la ré-indexation des documents. La taille du corpus affecte la durée de ré-indexation qui peut engendrer une lenteur du système, voire son indisponibilité pendant un certain temps. De même, la fréquence de mise à jour de la collection a un impact sur la disponibilité de l'index. En effet, plus la fréquence de mise à jour de la collection est élevée, moins l'index est disponible pour la RI car il est à tout moment en cours de modification.

Nous présentons nos solutions relatives à la dynamique de la RI sémantique dans la section 3.3.

Par ailleurs, parmi les éléments non négligeables, qui affectent la qualité de la réponse du système, au cours d'une session de RI figurent les contextes de la recherche que nous allons présenter dans la section suivante.

2.5. Contexte en Recherche d'Information

Toute activité de RI se déroule autour de trois éléments fondamentaux : l'utilisateur qui a besoin d'information, la tâche de recherche proprement dite et le corpus de documents dans lequel les documents potentiellement pertinents sont recherchés (Ingwersen et Järvelin, 2005). Ces trois éléments sont présentés successivement ci-après.

2.5.1. Utilisateur

L'utilisateur est l'acteur principal de toute activité de RI. C'est lui qui exprime ses besoins sous forme de requête soumise au SRI. C'est lui seul qui peut juger de la pertinence des résultats par rapport à ses attentes. Ainsi, pour bien satisfaire les besoins en information de chaque utilisateur, la prise en compte de l'utilisateur est très importante dans la RI. En effet, les besoins en information varient d'un utilisateur à l'autre, et ce en fonction de son profil. Par exemple sur un même SRI, les besoins en information varient selon que l'utilisateur soit enseignant ou apprenant, voyant ou malvoyant. A cela s'ajoute l'ensemble des connaissances pré-acquises par l'utilisateur.

Ainsi, pour être bien concordante, la restitution d'information à l'utilisateur doit être adaptée en conséquence.

2.5.2. Tâche

La tâche à réaliser dans un SRI est bien sûr la recherche de documents potentiellement pertinents par rapport aux besoins de l'utilisateur. Quand l'utilisateur a besoin d'apprendre ou d'approfondir une notion quelconque, la RI est déclenchée et le besoin en information est exprimé sous forme de requête soumise au SRI. Pour satisfaire les besoins de l'utilisateur, la tâche de recherche doit prendre en compte non seulement la requête exprimée par l'utilisateur, mais également les types de documents recherchés ainsi que les caractéristiques ou profils de l'utilisateur (Hernandez, 2005).

2.5.3. Document

Toute RI se résume en la recherche des documents contenant les informations recherchées. Comme l'objectif des recherches d'information est l'apprentissage de nouvelles notions, la présentation des documents contenant ces informations doit s'adapter aux contextes de l'utilisateur pour qu'il comprenne la sémantique de leurs contenus. Une description détaillée des modèles de document est présentée dans la section 2.1.2.

Dans le cadre de cette thèse, les documents sont des objets pédagogiques pour le système d'apprentissage en ligne, les partitions musicales Braille pour le système de recherche des documents Braille et des fiches de maintenance pour le SRI spécialisé dans l'aide au diagnostic.

2.5.4. Conclusion

La meilleure satisfaction des utilisateurs au cours d'une session de RI ne s'obtient qu'en tenant compte de ces trois éléments de base des SRI. Ainsi, dans le chapitreChapitre 3, nous présentons nos propositions théoriques pour la prise en compte de ces briques de base des SRI et nous développons dans le chapitreChapitre 4 les applications mettant en œuvre nos propositions complètes. Un cas particulier de RI est l'apprentissage en ligne qui est présenté dans la section suivante.

2.6. Cas de l'apprentissage en ligne

L'apprentissage en ligne est une activité pédagogique qui vise à acquérir ou à approfondir des connaissances tout en repoussant les contraintes de temps et d'espace entre l'apprenant et l'enseignant, par l'utilisation des nouvelles technologies de l'information et de la communication (E-tud), (Boutemedjet,2004).

2.6.1. Système d'apprentissage en ligne

Un système d'apprentissage en ligne doit permettre (Hernandez et al., 2006) :

- l'accès aux ressources pédagogiques pertinentes grâce à une bonne indexation des ressources (Gasevic et Hatala, 2005) (Psyché et al., 2005) (Lenne et al., 2005) (Abel et al., 2003),
- une interaction et une navigation suivant une pédagogie d'apprentissage adéquate mise en place (Psyché et al., 2005),
- la réutilisabilité des objets et des scenarii pédagogiques (Knight et al., 2005),
- la conception et la mise à jour du contenu des cours par les enseignants (Lenne et al., 2005) (Abel et al., 2003),
- le suivi individualisé des apprenants (IMSLD, 2003).

Les différents types de système d'apprentissage en ligne sont :

- le Content Management System (CMS) qui s'occupe de la gestion de contenu, c'est-à-dire la création et l'élaboration des objets pédagogiques,
- le Learning Management System (LMS) qui s'occupe de la gestion de l'apprentissage et de l'enseignement. L'administration et l'encadrement des participants y est fondamental,

- le Learning Content Management System (LCMS) qui comprend à la fois les fonctionnalités d'un CMS et d'un LMS. Il a donc comme fonctionnalité la gestion de contenu, la gestion de l'apprentissage et de l'enseignement.

2.6.2. Objet pédagogique

Les objets pédagogiques sont des documents électroniques créés dans l'objectif d'être intégrés dans un environnement technologique dédié à l'apprentissage en ligne. Ce type de documents reflète donc les enjeux majeurs de tout document électronique : utilisation dans différents scénarii d'usage (ici pédagogiques), réutilisabilité (utilisation de tout ou partie d'une ressource pour en construire une autre, pour d'autres objectifs ou d'autres utilisateurs), confrontation aux normes en cours d'élaboration et aux environnements technologiques qui les manipulent (ici les plates-formes d'apprentissage en ligne ou même Internet). Nous nous intéressons donc à ces documents électroniques spécifiques que sont les objets pédagogiques utilisés dans les environnements d'apprentissage en ligne.

Le groupe de travail IEEE-LTSC (Learning Technology Standards Committee) propose deux définitions d'un objet pédagogique :

- un objet pédagogique peut être défini comme « toute entité numérique ou non, qui peut être utilisée, réutilisée ou référencée lors d'une formation dispensée à partir d'un support technologique ». Cela permet de considérer comme objet pédagogique un document imprimé, un cours, un exercice, une étude de cas, une présentation (LOM, 2002),
- « un objet pédagogique est défini comme toute entité numérique ou non qui peut être utilisée, réutilisée ou référencée pendant des activités d'apprentissage assistées par ordinateur (enseignement intelligent assisté par ordinateur, environnements d'enseignement interactifs, systèmes d'enseignement à distance, environnements d'apprentissage collaboratif) » (LOM, 2002).

Dans nos travaux qui visent à l'indexation sémantique des objets pédagogiques par l'usage des métadonnées, nous adoptons plutôt cette deuxième définition.

Cependant, l'utilisation des métadonnées telle qu'elle est préconisée ne suffit pas et ne résout pas les problématiques des systèmes d'apprentissage en ligne : réutilisabilité et accessibilité.

Un problème complémentaire relève du fait que le système et les acteurs doivent partager le même sens accordé aux valeurs des métadonnées. D'autre part, les liens et relations comme la composition, l'ordre d'apprentissage, et les dépendances de pré-requis entre objets pédagogiques doivent être mentionnés pour permettre non seulement de réaliser des traitements ou tâches automatiques sur ces objets mais aussi de produire de nouvelles connaissances à partir de celles déjà existantes.

L'utilisation des ontologies dans le modèle sous-jacent à un système d'apprentissage en ligne nous ont paru une solution intéressante. En effet, comme nous le verrons dans la section suivante, une ontologie peut fournir le vocabulaire et sa sémantique, ainsi que la structure des métadonnées associées aux ressources annotées.

2.6.3. Les normes associées aux systèmes d'apprentissage en ligne

L'application des normes du domaine de la formation en ligne, qui sont considérées comme des langages communs de description des ressources éducatives numérisées (Vidal, 2004), garantit non seulement l'interopérabilité mais également la qualité du système. Parmi les normes de la formation en ligne, on peut citer SCORM, LOM et IMS-LD. LOM s'intéresse à la description des ressources pédagogiques, SCORM à la structure du contenu des objets, et IMS-LD au scénario d'apprentissage. Ces différentes normes sont détaillées en annexe.

2.6.3.1. LOM

Le standard LOM, spécifie la syntaxe et la sémantique des métadonnées décrivant des ressources pédagogiques numériques ou non et définit les attributs nécessaires à une description complète des ressources pédagogiques à partir des 78 attributs divisés en ces neuf catégories détaillées en annexe.

LOM permet d'identifier tout type de contenu, et plus spécifiquement les contenus éducatifs, avec un format commun décrivant par exemple le type de contenu, son auteur et la meilleure manière de l'utiliser.

Néanmoins, devant la difficulté de mise en œuvre du modèle LOM, l'utilisation des profils d'application devient nécessaire en vue de permettre d'affecter un caractère optionnel ou obligatoire aux champs des métadonnées.

2.6.3.2. SCORM

SCORM (Sharable Content Object Reference Model) (SCORM, 2004) d'Advanced Distributed Learning (ADL) est un modèle de référence pour le partage de contenus et d'objets. SCORM est un modèle pour l'assemblage des contenus web et un environnement d'apprentissage pour les objets pédagogiques. Il a pour vocation la mise en place de la bonne structuration du contenu du cours et de ses interactions avec son environnement.

La structuration du contenu des modules d'enseignement suivant le modèle SCORM permet de les réutiliser dans d'autres modules pour différentes formations ou systèmes. De plus, elle améliore le dialogue entre les objets pédagogiques et le système d'une part, et entre les acteurs et le système d'autre part.

Les objets pédagogiques répondant au modèle SCORM doivent avoir les caractéristiques RAID suivantes :

- Réutilisabilité : le contenu est indépendant du contexte d'apprentissage et peut être utilisé par plusieurs apprenants,
- Accessibilité : le contenu peut être identifié et localisé à tout moment,
- Interopérabilité : le contenu fonctionne dans tous les environnements (matériels et logiciels),
- Durabilité : le contenu ne nécessite pas de modification suite à un changement ou une mise à jour de système d'exploitation.

2.6.3.3. IMS-LD

Un Learning Design ou scénario pédagogique est défini comme une description d'une méthode permettant à un apprenant d'atteindre certains objectifs d'apprentissage par la réalisation de certaines activités pédagogiques ordonnées, dans un environnement d'apprentissage (IMS LD Specification).

En complément de SCORM, qui s'intéresse surtout aux pédagogies de transfert, IMS-LD (IMSLD03) ou Instructional Management System Learning Design est une norme qui vise à apporter des éléments de pédagogie dans un système d'apprentissage en ligne. Il s'agit d'un langage de modélisation des processus d'apprentissage, basé sur les travaux de Koper (Koper et al., 2001).

Il a été conçu pour la définition de scénarii d'apprentissage et d'interaction pour les créateurs de contenu ou de cours. Il aide les concepteurs à modéliser qui fait quoi, quand et avec quelles ressources et quels services pour réaliser des objectifs d'apprentissage.

Il définit la structure d'une unité d'apprentissage comme une « pièce » : c'est-à-dire un ensemble « d'actes » composés de « partitions » associant des « activités » à des « rôles » (enseignant, apprenant,…). Une représentation conceptuelle d'IMS-LD est présentée dans la Figure 57 en annexe. Pour faciliter la production de la spécification ainsi que son implémentation, IMS-LD a été divisé en trois parties :

- le niveau A contient l'ensemble des structures de base incluant : Activités, Environnements, Pièces, Actes, Rôles, Services,
- le niveau B ajoute des Propriétés et Conditions au niveau A. Ceci permet la personnalisation, le séquencement et l'interaction plus élaboré basés sur le profil de chaque apprenant,
- le niveau C ajoute les notifications au niveau B. Une notification est déclenchée par la réalisation d'un Résultat, et rend une activité exécutable et disponible pour un Rôle.

Chaque niveau est représenté dans des fichiers XML séparés. Les niveaux B et C intègrent et étendent le niveau précédent.

Dans notre modèle, nous nous appuyons sur IMS-LD pour définir le déroulement des interactions Homme-Machine pendant la phase d'exécution et d'utilisation des objets pédagogiques.

Les normes définies dans le contexte de l'apprentissage en ligne permettent de s'assurer d'une certaine interopérabilité et utilisabilité (au travers de scénarii).

2.6.4. Apprentissage en ligne et ontologie

Malgré la présence des normes telles que SCORM et LOM pour la description des ressources pédagogiques, l'application des normes n'est pas suffisante pour assurer la réutilisabilité et l'indexation des ressources. Toutefois, l'utilisation d'ontologies pour l'indexation des ressources permet d'enrichir ces différentes normes. Différents travaux de la littérature s'appuient sur des ontologies pour tenter de résoudre les problématiques que sont la réutilisabilité, l'accessibilité ainsi que l'indexation sémantique des contenus des ressources pédagogiques.

MEMORAE (pour MEMoire ORganisationnelle Appliquée à l'apprentissage en ligne) (Lenne et al., 2005) (Abel et al., 2003) est un outil d'apprentissage en ligne et d'indexation de ressources. Cet outil met à disposition des ressources pédagogiques aux apprenants, soit au sein d'une banque de ressources locales, soit dans un emplacement distant sur le Web, référencé par son URI. Memorae a pour objectif de faciliter l'autorégulation de l'apprentissage en explicitant les connaissances à appréhender ainsi que les relations qui existent entre elles et en leur associant des ressources appropriées.

Par rapport à MEMORAE qui présente des cours structurés suivant les relations d'inclusion, d'utilisation, de référence et de pré-requis entre les notions à appréhender, (Gasevic et Hatala, 2005) proposent, en plus, pour la recherche de ressources pédagogiques un outil permettant aux utilisateurs de formuler des requêtes libres, dans un champ prévu à cet effet. Cela permet ainsi aux utilisateurs de rechercher de l'information dans une banque de ressource distante. Par ailleurs, cet outil respecte la norme LOM quant à l'annotation et l'indexation des ressources pédagogiques. Ces deux études représentent la connaissance du système à l'aide d'ontologies. Une ontologie de domaine de la formation décrit les concepts tels que les personnes (étudiants, tuteurs, secrétaires…), les documents (livres, supports de présentation, pages web…). Elle est appelée ontologie du domaine de la formation pour (Lenne et al., 2005) et ontologie cible pour (Gasevic et Hatala, 2005). Une autre ontologie décrit les notions à appréhender, appelée ontologie d'application pour (Lenne et al., 2005) et ontologie source pour (Gasevic et Hatala, 2005).

(Hernandez, 2005) met l'accent sur la séparation des aspects de tâche et de thème tout en les mettant en relation. Chaque aspect est modélisé par une ontologie de domaine. L'ontologie de thème spécifie les notions qui doivent être assimilées par des étudiants pour une formation donnée. L'ontologie de tâche a pour but de préciser les contextes de l'apprentissage en spécifiant les ressources disponibles (ouvrage, logiciel,…), les modules qui composent ces ressources, leur type (cours, exercices, évaluation) ainsi que l'ordre dans lequel ils doivent être étudiés. Cette formalisation permet d'établir les connaissances associées à ces deux aspects à travers des relations sémantiquement riches. Le système d'apprentissage présente cette formalisation à l'utilisateur par un mécanisme d'exploration du corpus pédagogique reposant sur les deux ontologies. L'utilisateur appréhende ainsi le contexte associé aux objets pédagogiques.

Pour la recherche de ressources pédagogiques dans une banque de ressource distante, (Gasevic et Hatala, 2005) considèrent de plus une ontologie de correspondance. Cette dernière sert à décrire la correspondance ou la similarité entre les concepts de l'ontologie source qui prend en compte le contexte du cours présenté avec ceux de l'ontologie cible qui décrit la banque de ressource distante. Toutefois, l'inconvénient de ce système est que son efficacité dépend fortement de l'ontologie de mise en correspondance car une connaissance préalable de la structure de l'ontologie cible est nécessaire pour la mise en place de cette ontologie de correspondance. Par conséquent, ce système n'est pas facilement réutilisable car l'ontologie de correspondance doit être mise à jour à chaque modification de l'ontologie cible ou de l'ontologie source.

(Lenne et al., 2005) comme (Gasevic et Hatala, 2005) orientent leurs études autour de l'apprenant alors que (Knight et al., 2005) et (Psyché et al., 2005) considèrent aussi une assistance à l'auteur de la ressource pédagogique. La prise en compte de l'apprenant et de l'enseignant parait essentielle pour une utilisation optimale par ces deux types d'acteurs.

(Gasevic et Hatala, 2005), (Psyché et al., 2005) et (Knight et al., 2005) intègrent, contrairement à (Lenne et al., 2005), la notion de scénario pédagogique grâce à une ontologie basée sur la norme IMS-LD. Seul (Psyché et al., 2005) prennent en compte les théories éducatives afin de pallier au manque de relation entre le scénario d'apprentissage (Learning Design ou LD) et les théories des pédagogies d'apprentissage. Les différents types de théories de l'éducation sont représentés grâce à une troisième ontologie, appelée ontologie de la théorie éducationnelle.

Comme dans (Lenne et al., 2005) et (Hernandez, 2005), (Bouzeghoub et al., 2005) utilisent une ontologie de domaine qui représente l'ensemble des concepts décrivant les connaissances du domaine à appréhender en vue de l'indexation sémantique des ressources. Cependant

(Bouzeghoub et al., 2005) se distinguent par l'utilisation de deux types de métadonnées : l'un pour la description des caractères éducatifs et l'autre pour le caractère sémantique des ressources (Pré-requis, contenu et fonction pédagogique). De plus, afin de mieux décrire les ressources pédagogiques aussi bien que les apprenants, ils utilisent trois modèles: le modèle de domaine qui sert de référentiel pour indexer sémantiquement tant les apprenants que les ressources, le modèle de l'apprenant qui est utilisé dans le processus d'adaptation (filtrage) de ressources et enfin le modèle de description de ressources, qui permet de retrouver les ressources en vue de les réutiliser.

Dans toutes ces études ((Gasevic et Hatala, 2005), (Psyché et al., 2005), (Lenne et al., 2005)), le contexte d'utilisation des objets pédagogiques n'est pas pris en compte. Afin d'augmenter la réutilisabilité des scénarii et des objets pédagogiques, seuls (Knight et al., 2005) introduisent une ontologie de contexte. Ils associent aux objets pédagogiques (appelés LO pour Learning Object) un à plusieurs objets de contexte (appelés LOC pour Learning Object Context). Une séquence d'activités de l'apprenant est ainsi composée d'activités, une activité étant associée à des LO et des LOC.

2.6.5. Conclusion

L'apprentissage en ligne est un cas particulier de RI qui reflète bien les différentes propriétés des SRI dont les contextes ainsi que les dynamiques en RI que nous avons présentées dans les sections précédentes. Cependant, parmi les contextes qui n'ont pas été traités dans le cas de l'apprentissage en ligne figure le suivi des connaissances acquises par les utilisateurs afin non seulement de proposer des scénarii de lecture mais également de s'assurer de l'acquisition des connaissances. Un autre manque concerne la prise en compte des aspects pédagogiques pour le bon déroulement de l'apprentissage.

De nombreuses ressources pédagogiques sont accessibles via des moteurs de recherche sur Internet mais celles-ci correspondent souvent à de simples présentations en ligne de documents qui n'ont pas été créés spécifiquement pour leur exploitation dans des environnements d'apprentissage. Ce même problème se retrouve lorsqu'il s'agit de plates formes d'apprentissage qui deviennent des espaces organisés de ressources mais auxquelles ne sont pas rattachées de véritables situations d'utilisation. Ainsi, (Psyché et al., 2005) constatent l'insuffisance ou l'absence de l'application d'une approche pédagogique, que ce soit au niveau de la présentation des ressources pédagogiques ou du séquencement des activités d'apprentissage dans les outils actuels. Pourtant, parmi les solutions proposées par les organismes de normalisation, IMS-LD (IMSLD, 2003) qui est en charge de la pédagogie d'apprentissage et de son déroulement intègre la notion de scénario pédagogique. D'autres normes telles que SCORM (SCORM, 2004) et LOM (LOM, 2002) aident à l'homogénéisation des représentations de ce type de documents et facilitent l'interopérabilité. LOM rassemble les différentes métadonnées nécessaires pour la description des ressources pédagogiques mais n'inclut pas la représentation sémantique des contenus ; SCORM permet la structuration des contenus d'objets pédagogiques et leurs relations avec l'environnement d'utilisation.

Ces représentations ne sont pas suffisantes pour permettre et assurer la réutilisabilité de ressources ou de parties de ressources. La réutilisation de parties de ressources nécessite d'une part que la structure du document initial soit suffisamment marquée pour être exploitée, et d'autre part que le contenu sémantique ainsi que la portée d'usage soient suffisamment explicites pour chacune des parties.

Pour résoudre ces problèmes, le modèle général de représentation multi-facette présenté dans la section 3.1 est instancié pour les cas des objets pédagogiques dans la section 3.1.2. Un prototype de ce système est présenté dans la section 4.1.

2.7. Bilan

L'efficacité des SRI dépend de leur capacité d'analyse des contenus des documents. Cependant, divers types de documents se trouvent sur la toile. Plusieurs travaux ont été élaborés dans le but de permettre la recherche sémantique sur ces documents. Ces travaux s'étendent dès la modélisation des documents jusqu'à la recherche des documents en passant par les modèles de recherche et les fonctions de similarité entre les requêtes et les documents.

Quant à la modélisation des documents, (Pogodalla, 2004) et (Roisin, 1998) ont développé des modèles conceptuels de documents pour pouvoir comprendre et analyser leurs contenus. Cependant il manque d'une part la considération de l'utilisateur et du contexte dans lequel il veut utiliser le document et d'autre part la description sémantique du contenu des documents afin de mieux comprendre les messages véhiculés dans les documents.

Ceci nous conduit à proposer un modèle générique de représentation multi-facette des documents afin de prendre en compte les différents paramètres qui influencent les sémantiques des documents que nous avons ensuite instancié dans différents cadres.

Nous avons analysé dans le cadre d'un projet Européen un cas particulier de documents que sont les partitions musicales. Des formats d'encodage (MusicXML, ni NIFFML, ni MIDI) issus des normes et standards (SMDL, SMR) ont été élaborés pour encoder les partitions musicales, cependant aucun de ces formats n'a pris en compte la partition musicale Braille. Ceci nous amène à instancier notre modèle de représentation multi-facette des documents dans le domaine de la musique Braille et de proposer en même temps un format d'encodage des partitions musicales Braille qui tienne compte non seulement des normes sur les documents musicaux mais également des particularités de la musique Braille et ses contextes d'utilisations.

Les objets pédagogiques constituent un autre cas d'étude. Ces objets sont manipulés par les systèmes d'apprentissage en ligne. Il existe des normes comme SCORM, LOM, IMS-LD qui visent respectivement à leur structuration, à la description de leurs contenus et à leur séquencement. Cependant, il manque la notion d'approche pédagogique pour l'apprentissage non seulement au niveau de la présentation des ressources mais également au séquencement de l'apprentissage. Pour pallier ce manque, nous avons instancié notre modèle de représentation multi-facette de documents dans le domaine de l'apprentissage en ligne tout en proposant la prise en compte de l'approche pédagogique (cf. section 3.1.2).

L'un des objectifs de la modélisation des documents est de faciliter les recherches ultérieures de ces documents. Divers approches ont été élaborées pour retrouver les documents. Cependant, les modèles actuels de RI sont trop formels, et présupposent connus soit la fonction et soit les variables qui fondent la classification (Rousseaux et Bonardi, 2007). Parmi ces approches se trouvent la RI sémantique qui est le cadre de notre recherche. Dans ce type de RI, les documents sont annotés à l'aide des concepts issus d'une ontologie de référence. Pour retrouver ces documents, les SRI calculent les similarités entre les concepts annotant les documents avec ceux de la requête. Diverses mesures de similarités conceptuelles ont été développées (Wu et Palmer, 1994), (Resnik, 1995), (Jiang et Conrath, 1997) (Lin, 1998) (Leacock et Chodrow, 1998). Néanmoins, il manque une fonction de similarité sémantique générique pour permettre la comparaison entre les documents et les requêtes, et d'autre part

une mesure de similarité conceptuelle basée sur la sémantique des concepts et adaptée à divers types d'application. Ainsi, nous avons proposé des fonctions de similarité conceptuelle qui tiennent compte de la hiérarchie des concepts qui reflètent les proximités sémantiques entre concepts.

Les SRI sont, en général, destinés à être utilisés sur la toile ou sur des collections de documents dynamiques. Divers travaux ont été élaborés afin de prendre en compte les dynamiques des documents et de gérer leurs impacts sur les index (Ukkonen, 1985) (Cutting et Pedersen 1990) (Salton et al., 1993) (Tomasic et al., 1994) (Page et Brin, 1998) (Cho et Garcia-Moulina, 2000) (Galambos, 2006) (Büttcher et Clarke, 2006) (Lim et al., 2007). Des problématiques comme la disponibilité des documents et de mise à jour des index restent à résoudre. Pour résoudre ce problème, nous avons conçu un modèle d'indexation sémantique à base d'ontologies ainsi que les algorithmes de mise à jour correspondants.

L'ensemble des contributions sont décrites dans le chapitre 3.

> Man, he is constantly growing and when he is bound by a set pattern of ideas or way of doing things, that's when he stops growing.
>
> [Jun Fan]

Chapitre 3. PROPOSITIONS

L'assimilation des informations véhiculées dans les documents dépend de plusieurs paramètres. Parmi les éléments qui influencent la sémantique des informations nous pouvons citer la structuration logique et physique du document, son contexte d'usage ainsi que la description et annotation associée non seulement au document mais également aux différentes parties du document. En effet, le contenu d'un document quelconque peut prendre des sémantiques différentes selon la tâche que l'on veut exécuter sur ce document dans un contexte donné. Par exemple, dans un contexte où l'on s'intéresse à l'histoire, une image d'un palais royal peut être considérée comme un monument historique tandis que dans un contexte où l'on s'intéresse à la construction, c'est plutôt les informations relatives à l'architecture et aux matériaux utilisés qui peuvent y être considérées.

Ainsi, pour permettre une recherche sémantique des documents, il peut s'avérer utile de les représenter sous différents points de vue. Dans notre modèle, nous décrivons non seulement le contenu des documents, mais également leurs contextes d'utilisation afin de faciliter leurs recherches ultérieures.

Dans la section 3.1, nous présentons le modèle générique de représentation multi-facette de documents que nous proposons, ainsi que ses instanciations dans trois cadres applicatifs. Dans la section 3.2, nous décrivons l'indexation sémantique des documents. La section 3.3 s'intéresse à la dynamique de cette indexation lors des modifications de la collection ou des documents qui la composent. Dans la section 3.4 nous décrivons les principes de recherche de documents que nous proposons.

3.1. Représentation multi-facette de documents

Nous présentons ci-après un méta-modèle de représentation multi-facette des documents en vue de leur recherche ou réutilisation. Ce méta-modèle a été conçu de manière générique pour être instanciable dans divers cas d'utilisation. Nous présentons ensuite l'instanciation de ce méta-modèle dans plusieurs applications.

La section 3.1.1 présente le méta-modèle générique de représentation multi-facette. Les sections 3.1.2 à 3.1.4 présentent les instanciations de ce modèle respectivement dans le cadre des objets pédagogiques, des documents de maintenance et des documents Braille.

Ces différents modèles sont tous issus d'une instanciation en totalité ou en partie du méta-modèle, selon les besoins du domaine d'application.

3.1.1. Méta-modèle conceptuel de représentation multi-facette de documents

Nous représentons dans la figure 2, le modèle de données du système d'information. Ce modèle comprend les différentes classes permettant la représentation multi-facette des documents d'une collection quelconque, quel que soit le contexte d'application.

Ce méta-modèle est composé de trois parties inter-reliées dont :

- la composition des documents,
- la description des documents par des facettes,
- la tâche associée au document dans son contexte d'usage.

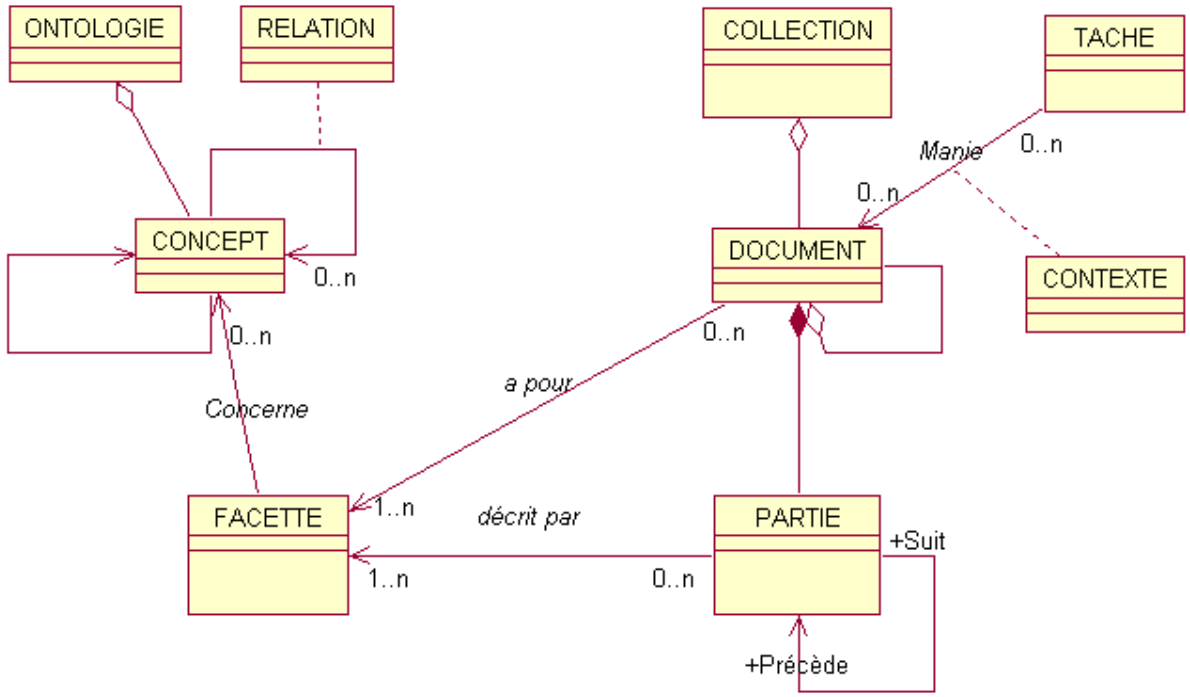

Figure 2 : Méta-modèle conceptuel de représentation multi-facette de documents.

Nous présentons ci-après les différentes parties qui composent le méta-modèle conceptuel de représentation multi-facette des documents avant d'instancier ce méta-modèle dans des applications concrètes.

3.1.1.1. Collection de documents et structure des documents.

La classe *COLLECTION* représente les collections de documents (corpus) qui rassemblent les documents relatifs à des domaines d'application. Une collection est composée d'un ou plusieurs documents. Chaque document peut être découpé en plusieurs parties. Ces liens entre documents et parties de documents peuvent être une relation de composition physique et/ou une répartition logique. Chaque partie de document est en relation avec celles qui la précède ou la succède dans l'enchaînement logique de lecture dudit document.

3.1.1.2. Représentation des facettes

Une ontologie, qui est une représentation formelle et partagée des connaissances d'un domaine, est composée de concepts et de relations entre ces concepts. Une relation peut être un lien d'héritage (hiérarchie de concepts ou subsomption) ou bien une relation sémantique entre concepts («a pour » entre DOCUMENT et FACETTE, « décrit par» entre PARTIE et FACETTE). La classe *FACETTE* représente l'ensemble des métadonnées décrivant les documents ou parties de documents.

Les contenus des documents sont annotés à l'aide des métadonnées. Ces dernières servent à décrire les documents et les parties des documents et prennent leurs valeurs à partir des termes décrivant les concepts de l'ontologie de domaine. L'indexation sémantique des contenus des documents, utilisant les valeurs de ces métadonnées, est traitée en détails à la section 3.2.

3.1.1.3. Description de la tâche d'usage des documents

Chaque document peut être utilisé dans des tâches relatives à son domaine d'utilisation. Ces tâches sont représentées par la classe *TACHE*. Un même document peut être utilisé dans plusieurs tâches dans un contexte bien déterminé. Un contexte de recherche de documents met en relation plusieurs entités dont les utilisateurs, les documents, ainsi que les tâches associées à ces documents. Ainsi, un contexte dans le cadre de la RI peut être défini comme une utilisation particulière d'un document par un utilisateur dans une tâche.

Une instanciation donnée du méta-modèle n'instancie pas forcément toutes les classes ou aspects présents dans le méta-modèle. Dans ces conditions, les trois aspects (documents, facettes, tâches) ne se retrouvent pas forcément dans toutes les applications ; les facettes dépendent également des applications. Une des facettes récurrentes correspond à la description thématique des documents via une ontologie. L'association entre les documents et l'ontologie qui est réalisée lors de l'indexation est présentée dans la section 3.2.

Dans la section suivante, nous présentons différentes instanciations de notre méta-modèle dans différentes applications avec différents types de documents dont les objets pédagogiques, les documents de maintenance automobile et les documents musicaux Braille.

3.1.2. Modèle de représentation multi-facette des objets pédagogiques

Dans le cadre du domaine de l'apprentissage en ligne, le méta-modèle que nous avons spécifié est instancié afin d'aboutir à une représentation multi-facette des objets pédagogiques (cf Figure 3) que nous décrivons ci-après.

3.1.2.1. Représentation multi-facette des objets pédagogiques

Afin d'une part de disposer d'un système d'apprentissage qui utilise des approches pédagogiques adéquates pour mieux apprendre les notions et connaissances relatives à un domaine d'étude particulier, et d'autre part de permettre la réutilisabilité des objets pédagogiques et des scénarii pédagogiques (Learning Design ou LD), nous proposons de modéliser les différents aspects permettant de décrire les objets pédagogiques. Comme notre méta-modèle le permet, nous distinguons la description propre à l'objet pédagogique et celle liée à ses usages.

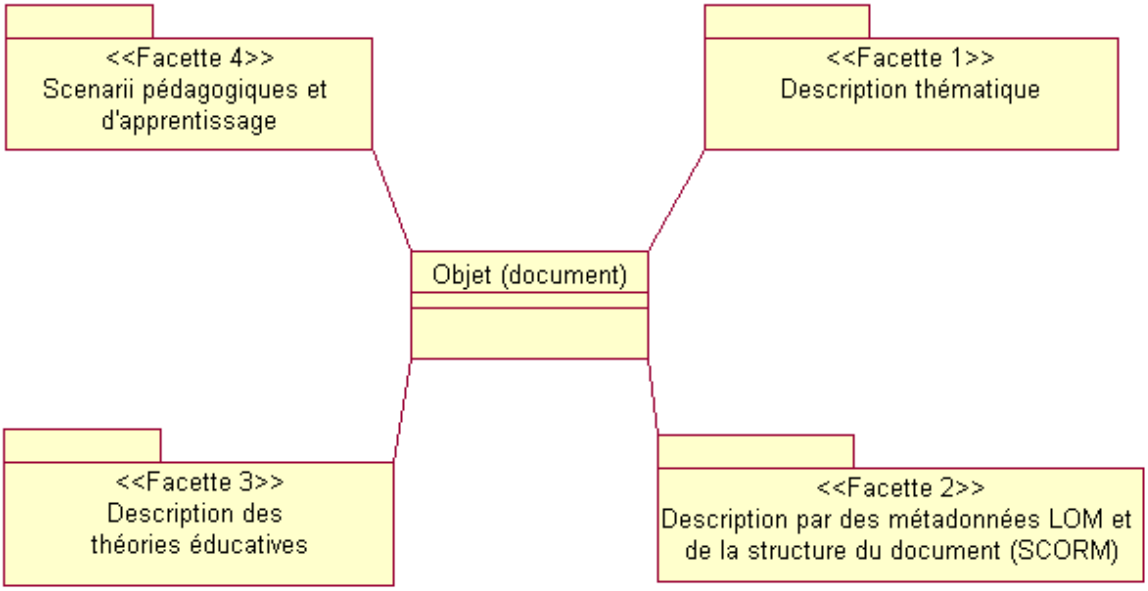

Figure 3 : Connaissances utiles pour représenter un document (objet pédagogique) et son usage (Hernandez et al., 2006).

Les documents pédagogiques sont composés d'objets pédagogiques, c'est-à-dire d'agrégats de contenu et objet de contenu partageable (SCO pour Shareable Content Object) et de composants élémentaires (Assets). Ils abordent des notions d'un domaine donné et sont inclus dans des scenarii pédagogiques. Pour représenter un document pédagogique nous considérons différentes connaissances :

- connaissances sur la ressource elle-même (norme LOM) et sur la structuration du document (norme SCORM pour Shareable Content Object Reference Model),
- connaissances sur le thème abordé par le document,
- connaissances sur l'ensemble des théories éducatives existantes,
- connaissances sur le scénario pédagogique (norme IMS-LD).

Afin de représenter ces différents types de connaissances, nous proposons d'utiliser des ontologies. L'intérêt d'utiliser des ontologies dans ce contexte réside dans une représentation non ambigüe de la connaissance (en particulier la levée des ambiguïtés terminologiques). En associant les concepts des ontologies aux documents pédagogiques ou aux usages de ces documents (scénarii d'apprentissage), il est également possible d'induire un raisonnement grâce aux axiomes associés à celles-ci.

La facette « Métadonnées LOM et structuration SCORM» permet d'un côté de décrire et d'indexer tout objet pédagogique à l'aide des métadonnées de LOM d'un autre côté de structurer chaque objet pédagogique suivant la norme SCORM (cf. section. 2.6.3.2).

3.1.2.2. Structure des objets pédagogiques (SCORM)

Un document pédagogique (ou objet pédagogique) est une unité sémantique de ressource d'apprentissage. Il peut être un exercice, un sujet d'examen, une définition, des exemples, ou bien une leçon. Chaque objet pédagogique peut rassembler des composants élémentaires (comme une image) nommés Composant (appelé « asset » dans la norme SCORM) qui

peuvent être de formats numériques (.DOC, .PDF, .JPG) ou physiques différents. Un objet pédagogique peut par ailleurs être composé d'autres objets pédagogiques.

3.1.2.3. Description par des métadonnées (LOM et Profil d'application)

La description des métadonnées associées à un document pédagogique correspond à celle qui est prévue par LOM. Comme dans (Duval et al., 2002), nous proposons l'utilisation d'un *Profil d'application* qui est un assemblage d'éléments de métadonnées en vue d'adapter des schémas existants pour constituer un ensemble taillé à la mesure des exigences fonctionnelles d'une application particulière, tout en restant interopérable avec les schémas d'origine.

Dans notre proposition, une description LOM est rattachée à chaque objet pédagogique (qu'il soit élémentaire ou composé).

Les métadonnées utiles pour une application donnée peuvent donc ensuite être filtrées via le profil d'application, en fonction de celle-ci. Une modélisation de cette description est montrée dans la Figure 4. Lorsqu'un objet est utilisé dans une formation donnée, certaines valeurs des métadonnées associées à la formation elle-même sont automatiquement renseignées pour les objets pédagogiques associés.

De notre point de vue, ces informations ne correspondent pas à une connaissance nécessitant leur modélisation à travers une ontologie. Aussi, ces informations sont-elles simplement rattachées à la formation d'une part et aux objets pédagogiques d'autre part.

LOM est destiné à l'annotation et l'indexation des ressources pédagogiques. Les métadonnées associées permettent de renseigner, d'une manière normalisée, les différentes informations nécessaires sur chaque objet d'apprentissage, de façon à ce que les recherches ultérieures soient rendues plus efficaces.

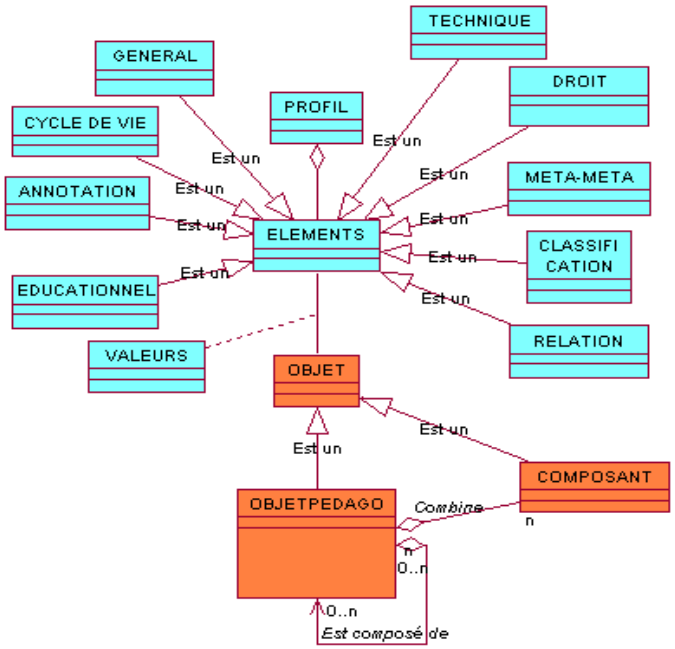

Figure 4 : Facette de Description par des métadonnées LOM et de la structure du document (SCORM) (Hernandez et al., 2008).

Dans cette figure, la classe OBJET est une instance de la classe DOCUMENT du méta-modèle que nous avons présenté à la page 42. De même, la classe ELEMENTS n'est autre qu'une instance de la classe CONCEPT de ce même modèle.

Dans notre modèle, nous utilisons SCORM pour représenter les structures des modules d'enseignement d'une formation et ainsi garantir leur interopérabilité. L'agrégat de contenu ainsi que l'objet de contenu partageable de SCORM sont tous représentés à l'aide d'un « objet pédagogique » tandis que la ressource numérique élémentaire est représentée à l'aide de « composant ».

3.1.2.4. Description thématique

Les objets pédagogiques sont également représentés par rapport aux thématiques ou notions qu'ils abordent. Dans le but de pouvoir réutiliser des documents ou des parties de documents abordant une notion traitée dans le cadre de plusieurs formations ou de plusieurs modules, les objets pédagogiques sont indexés à partir des concepts d'une ontologie de domaine du thème décrivant les thématiques abordées. Cette ontologie décrit l'ensemble des notions en lien avec le domaine et les représentent à partir de leur lien sémantique. Par exemple, dans l'ontologie du domaine des bases de données (cf Figure 26), la notion (concept) de « *base de données relationnelles* » se *« conceptualise »* à partir d'un « *modèle Entité-Association* ».

L'apprentissage d'une notion pouvant demander des pré-requis, les notions correspondant à des pré-requis d'une autre notion sont également représentées dans l'ontologie (si l'on reprend l'exemple précédent, la notion *d'attribut* doit être assimilée pour appréhender la notion de dépendance fonctionnelle). La Figure 5 montre un modèle de description d'ontologie de thème.

La représentation sémantique du contenu des objets pédagogiques à l'aide des métadonnées qui prennent leurs valeurs à partir des concepts d'une ontologie du thème présente différents avantages d'utilisation aussi bien pour l'enseignant que pour les apprenants. Ainsi, pour un module donné (par exemple un module de bases de données), les notions à assimiler sont précisées dans l'ontologie de ce thème et les objets pédagogiques relatifs à ce domaine sont indexés à l'aide des concepts de cette ontologie. Lorsqu'un enseignant souhaite concevoir une leçon, il peut ainsi avoir accès à l'ensemble des objets pédagogiques qui ont été indexés à partir des notions spécifiées pour le module. Il peut alors réutiliser les objets ou décider d'en concevoir de nouveaux s'ils ne lui conviennent pas. Du point de vue de l'apprenant, l'accès aux différentes notions en lien avec la formation et le module suivis lui permettent de situer ses connaissances (acquises ou à acquérir) dans le contexte d'apprentissage.

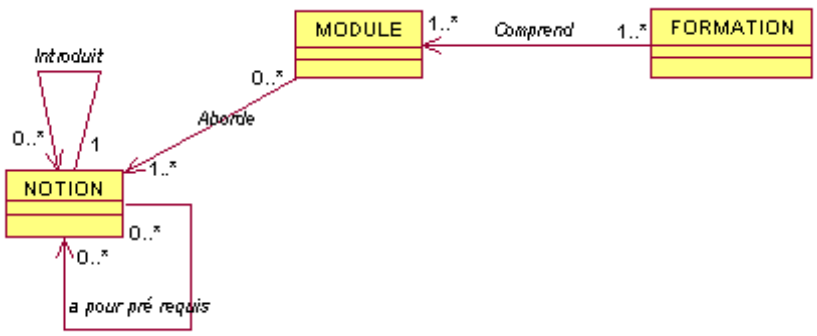

Figure 5 : Facette de description thématique (Hernandez et al., 2008)

3.1.2.5. Suivi des connaissances des utilisateurs

Les connaissances des utilisateurs sont exprimées à l'aide des concepts issus de l'ontologie de domaine. A l'issue de tous les cycles d'apprentissage, l'ontologie qui représente l'ensemble des connaissances des utilisateurs devrait être similaire à l'ontologie de domaine de référence.

En tant que représentation conceptuelle des connaissances du domaine de l'étude, l'ontologie de domaine sert de base de référence des termes servant à l'indexation des objets pédagogiques et des cours (objets pédagogiques avec scénario IMS-LD). Ainsi, l'ontologie de domaine est considérée comme un ensemble des connaissances qui devraient être acquises à l'issue des scénarii d'apprentissage utilisant des objets pédagogiques indexés à l'aide des termes des concepts de l'ontologie.

3.1.2.6. Ontologie personnelle des utilisateurs

L'ontologie personnelle, comme son nom l'indique, représente donc le profil ou l'ensemble des connaissances acquises par chaque apprenant à un moment donné. Elle reflète d'une part l'état d'avancement de chaque apprenant par rapport aux scénarii définis dans chaque cours et d'autre part le niveau de compréhension des notions abordées dans chaque activité d'apprentissage. De ce fait, l'analyse de l'ontologie personnelle permettra, en cas de besoin, de proposer des séquences d'activités supplémentaires (des activités d'apprentissage utilisant des objets pédagogiques qui traitent des notions que l'apprenant devrait maîtriser avant d'aborder un objet pédagogique donné) à un apprenant moyennant les relations « a pour prérequis » entre chaque objet pédagogique.

3.1.2.7. Validation des connaissances

Dans une activité d'apprentissage donnée, le passage à l'activité suivante est soumis soit à une validation par l'apprenant de sa compréhension et maîtrise des notions abordées dans l'activité courante, soit à une évaluation de l'étudiant à l'aide des exercices d'évaluation. Cette validation a pour effet de récupérer les valeurs des métadonnées des objets pédagogiques utilisés dans l'activité en cours et de les intégrer dans l'ontologie personnelle de l'apprenant.

3.1.2.8. Dynamique des ontologies personnelles

L'ontologie personnelle des utilisateurs apprenants se construit au fur et à mesure de leur avancement dans l'activité d'apprentissage. En effet, des objets pédagogiques, qui sont indexés avec le concept *Notion* de l'ontologie de domaine, sont associés à chaque activité d'apprentissage. Ainsi, à chaque réalisation d'une activité donnée par un apprenant, les différentes notions associées aux objets pédagogiques de cette activité peuvent être intégrées dans son ontologie personnelle.

Après avoir effectué toutes les activités pédagogiques de tous les cours, les ontologies personnelles des apprenants qui ont réussi leurs études avec succès devraient couvrir l'étendue de la partie de l'ontologie dont les termes sont utilisés dans l'indexation.

3.1.2.9. Usage dans les scenarii d'apprentissage (IMS-LD)

IMS-LD propose de modéliser la séquence des activités d'apprentissage attribuées à chaque rôle pour que l'objectif visé par l'apprentissage soit réalisé, tout en suivant une pédagogie bien déterminée. Les connaissances nécessaires pour prendre en compte les scenarii d'apprentissage sont les suivantes :

- connaissances sur l'ensemble de tous les intervenants ou acteurs qui participent à l'aboutissement d'une formation donnée. Les intervenants sont représentés par le *Rôle* dans notre modèle. Un rôle peut être « Enseignant », « Apprenant », « Tuteur » ou « Administratif ». A chaque rôle est associé un ensemble d'activités à réaliser,
- connaissance sur le déroulement de l'apprentissage d'un cours dans lequel le document est utilisé (scénario). IMS-LD l'appelle *Méthode*, il peut contenir une ou plusieurs *pièces*. Une pièce est composée d'*Acte*s qui sont exécutés séquentiellement. Les actes sont composés de *Partitions* qui associent un rôle à une activité effectuée dans un *Environnement* composé d'objets pédagogiques et de services (chat, forum, supports de cours…),
- connaissance sur les activités dans lesquelles le document est utilisé. Dans notre modèle, l'*Activité* décrit les tâches interactives qui se déroulent entre les différents acteurs à travers le système pour l'apprentissage d'une notion donnée. Une activité peut être une lecture d'une ressource pédagogique, un test, une simulation, une auto-évaluation, un exercice, un dialogue ou interaction directe entre apprenant et tuteur... Elle traite un ensemble de notions et de compétences,
- connaissance sur le *Contexte d'utilisation* de l'objet pédagogique : la réalisation d'une activité peut utiliser ou manipuler des objets pédagogiques comme support ou référentiel dans un contexte d'utilisation donné. Ainsi, un même objet pédagogique peut être considéré ou valorisé différemment d'une activité (d'une formation) à l'autre. Le *Contexte* nous permet de décrire l'usage de l'objet pédagogique dans l'activité.

L'ensemble de ces connaissances est représenté grâce à une ontologie, comme le montre la Figure 8. Des relations entre concepts sont introduites. Par exemple, le concept *Pédagogie* de l'ontologie des théories éducatives est relié au concept *Méthode* de l'ontologie du scénario pédagogique. Cela permet de guider l'auteur dans la conception du document suivant la pédagogie qu'il a choisie.

De même, le concept *Notion* de l'ontologie du domaine est relié au concept *Activité* car l'apprentissage d'une notion peut se réaliser dans une ou plusieurs activités. Cela nous permet de prévoir la réutilisabilité de la ressource. Un même objet peut être utilisé pour plusieurs notions et dans plusieurs activités. On exprime ainsi la réutilisabilité des objets pédagogiques. C'est l'agencement séquentiel des objets pédagogiques dans différentes activités qui assure la conformité avec le scénario pédagogique choisi.

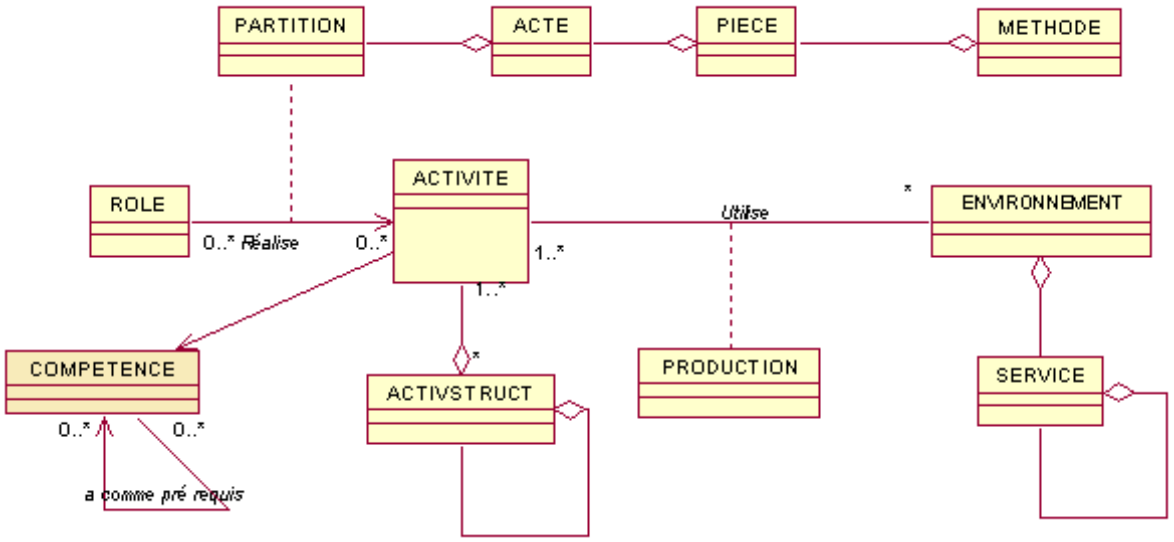

Figure 6 : Facette des scenarii pédagogiques et apprentissage

3.1.2.10. Description des théories éducatives

Selon les théories d'apprentissages suivies, chaque pédagogie appartient à un *Type* d'approche pédagogique (Empiriste, Rationaliste, Interactionniste) (Lebrun M., 2002) et est normalement constituée de plusieurs *étapes* distinctes à suivre.

Différents types d'approches pédagogiques existent :

- empiriste : pour l'empiriste, comprendre une réalité donnée, c'est avant tout savoir de quoi elle est faite, quels sont les faits qui la constituent,
- rationaliste : pour le rationaliste, comprendre une réalité donnée, c'est saisir la loi d'organisation de cette réalité, sa structure, abstraction faite du contenu particulier des faits,
- interactionniste : l'apprentissage est fondamentalement abordé comme le processus par lequel le savoir circule, se construit et se transforme au sein d'une communauté, d'un groupe social. Dans cette perspective, apprendre, pour l'individu, c'est participer à ce processus collectif de co-construction du savoir.

Une pédagogie choisie pourra donner lieu à plusieurs scenarii pédagogiques (*Méthodes).*

Une *Etape* désigne le découpage théorique d'une approche pédagogique donnée, comme illustrée dans la Figure 7. Une étape peut être une phase d'information, de motivation, d'interaction, de production, d'analyse… Une étape est associée à plusieurs *actes* dans le scénario pédagogique. En ce qui concerne l'ontologie des théories pédagogiques, le concept *Pédagogie* décrit l'ensemble des théories d'apprentissage qui peuvent être utilisées pour bien mener des formations.

Les connaissances associées aux théories pédagogiques sont représentées sous forme d'une ontologie. Cette représentation se justifie par le fait que nous souhaitons pouvoir associer des aspects raisonnements. Plus spécifiquement, l'aide à la construction d'un scénario à partir d'objets pédagogiques sera guidée par la connaissance préalable de la théorie d'apprentissage sous-jacente. Bien que les objets pédagogiques ne soient pas directement représentés à partir de cette ontologie, elle influence leur intégration dans le scénario pédagogique.

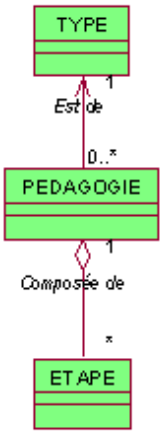

Figure 7 : Facette des théories éducatives (Hernandez et al., 2008)

3.1.2.11. Représentation multi-facette des objets pédagogiques

Ainsi notre modèle se base sur une représentation multi-facette des documents dont :

- une ontologie de thème (Hernandez, 2005) de la formation qui rassemble les thèmes, notions et connaissances à appréhender,
- une ontologie des tâches qui décrit les différentes activités d'apprentissage et d'enseignements, les organisations mises en place ainsi que les objets pédagogiques utilisés. Elle a été conçue dans le respect de la norme IMS-LD,
- une ontologie des théories pédagogiques qui décrit l'ensemble des différentes approches pédagogiques existantes. Elle est inspirée d'EML-OUNL (Koper, 2001),
- une description LOM qui permet de créer des profils d'applications pour la description des métadonnées utilisées aussi bien pour l'annotation des objets pédagogiques que pour la recherche de ces dernières, en réponse à des requêtes utilisateurs et/ou celles du système.

En utilisant ces quatre facettes tout en respectant les normes relatives à l'apprentissage en ligne que nous avons citées plus haut, nous obtenons le modèle de l'application présenté dans la Figure 8.

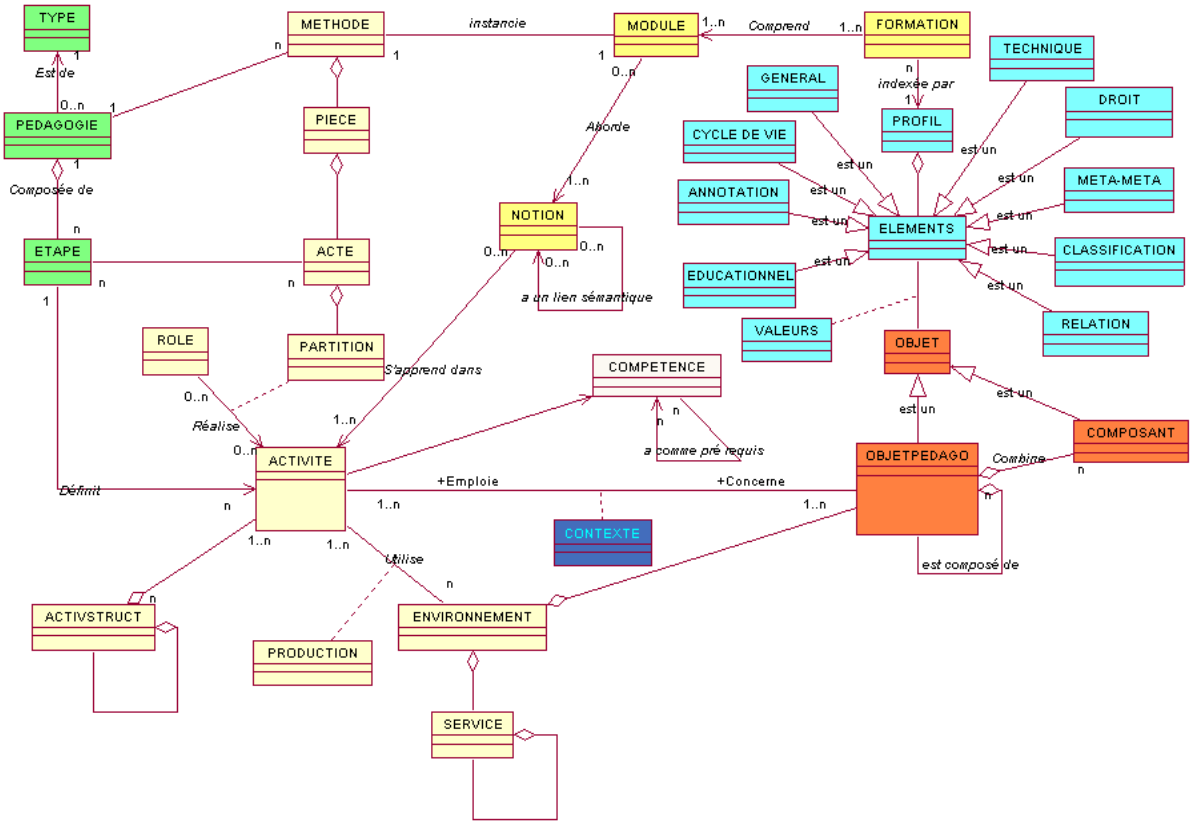

Figure 8 : Modèle complet intégrant les différentes facettes de description d'un document dans son contexte d'utilisation (Hernandez et al., 2008).

Le modèle que nous proposons traite tous les aspects du contexte d'un objet pédagogique :

- le niveau dans lequel l'objet est abordé est mentionné grâce aux métadonnées de LOM,
- son utilisation dans différents domaines spécifiques se traduit grâce aux différentes formations,
- les différents pré-requis sont considérés grâce aux notions pré-requises et aux compétences pré-requises,
- son usage dans différentes activités est précisé dans la classe contexte.

Le modèle proposé améliore la réutilisabilité d'un objet en considérant ses différents aspects : sa description par des métadonnées, son usage dans les scenarii d'apprentissage, son découpage et sa représentation sémantique.

Il facilite la recherche des objets pédagogiques, la gestion des droits d'accès aux ressources. Enfin, l'interopérabilité et la pérennité sont assurées grâce à la conformité du modèle avec les normes.

Ces travaux ont été publiés dans (Hernandez et al., 2008).

3.1.3. Modèle de représentation des documents de maintenance automobile

Ce travail a été réalisé dans le cadre du projet Dynamo (Dynamic Ontology for information retrieval).

L'objectif de la maintenance automobile consiste à maintenir les voitures en bon état de marche. Dans ce domaine, afin de mieux garantir la qualité de leurs prestations, les garagistes se servent des documents de maintenance délivrés par les constructeurs ou des experts en maintenance automobile.

Une fiche de maintenance est un document qui, pour un symptôme donné et pour chaque type de voiture, donne les informations relatives à la cause, aux solutions à appliquer et à la démarche de maintenance à suivre non seulement pour résoudre le problème mais également pour éviter que le problème ne se reproduise. Les garagistes sont donc amenés à rechercher, parmi les documents de maintenance présents dans le corpus, les bons documents relatifs à un symptôme constaté au moment du diagnostic. Ainsi, il leur faut un outil de recherche sémantique pour accéder aux bons documents. Pour arriver à cette fin, notre méta-modèle a été instancié dans le domaine de la maintenance automobile. Le modèle décrit les documents de maintenance automobile en vue de faire ressortir les sémantiques véhiculées dans les documents.

L'instanciation de chaque partie du méta-modèle donne lieu aux différents sous-modèles que nous allons présenter successivement. Nous y trouvons, les différentes représentations via les facettes structure, l'usage de ces documents ainsi que l'ontologie de domaine qui sert à décrire les connaissances du domaine.

3.1.3.1. Représentation des facettes des documents de maintenance

Pour permettre la description sémantique des contenus des documents de maintenances, nous considérons toutes les facettes qui permettent de décrire ces documents. Une facette de description thématique correspondant à l'ontologie du domaine du thème permet de décrire les connaissances nécessaires dans le domaine de la maintenance automobile. Une autre facette concerne la structure de ces documents. En effet, l'indexation de ce type de document ne se fait pas sur tout le contenu du document mais sur une partie seulement. Enfin, comme chaque fiche de maintenance peut être utilisée dans différents contextes d'usage, nous considérons aussi cette facette qui décrit les différents niveaux de connaissance en matière de maintenance automobile, suivant le contexte.

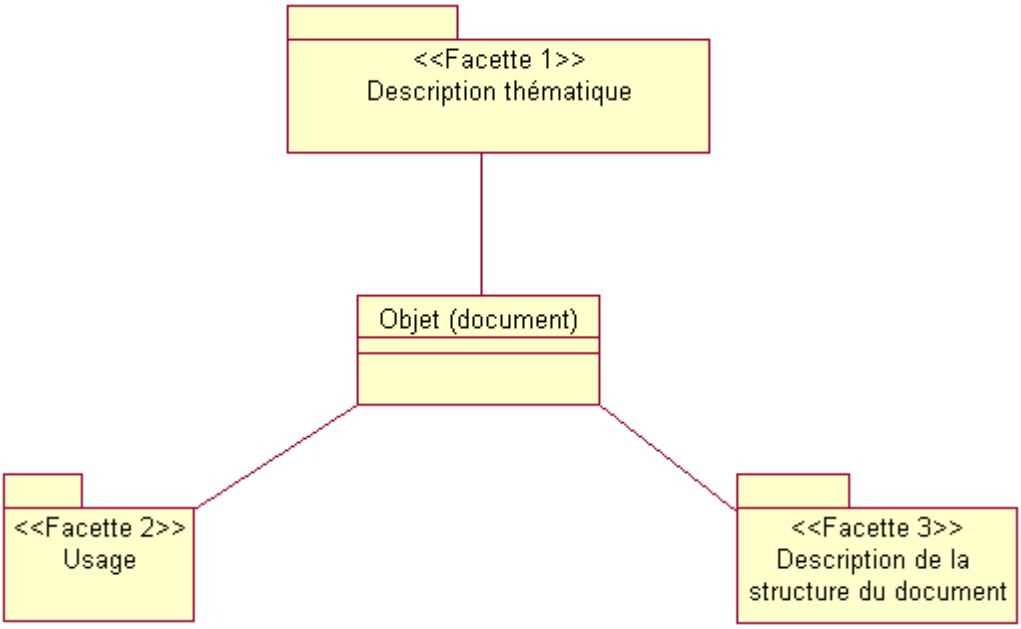

Figure 9 : Connaissances utiles pour représenter un document de maintenance et son usage

3.1.3.1. Structure des documents de maintenance automobile

Chaque fiche de maintenance (document) est découpée en cinq parties distinctes : symptômes, le défaut, le remède, le diagnostic et le type de voiture (cf Figure 10). Ainsi, pour un TYPE DE VOITURE donné, plusieurs fiches de maintenance se trouvent dans le corpus. Une fiche correspond à un SYMPTOME donné d'un type de voiture. La partie DEFAUT décrit les causes éventuelles du symptôme. La classe DIAGNOSTIC à son tour donne des informations plus détaillées concernant le fonctionnement des composants impliqués dans les symptômes. La classe REMEDE, affiche les différentes étapes à suivre pour traiter les problèmes entraînant les symptômes.

Nous présentons ci-après la structure des fiches de maintenance automobile.

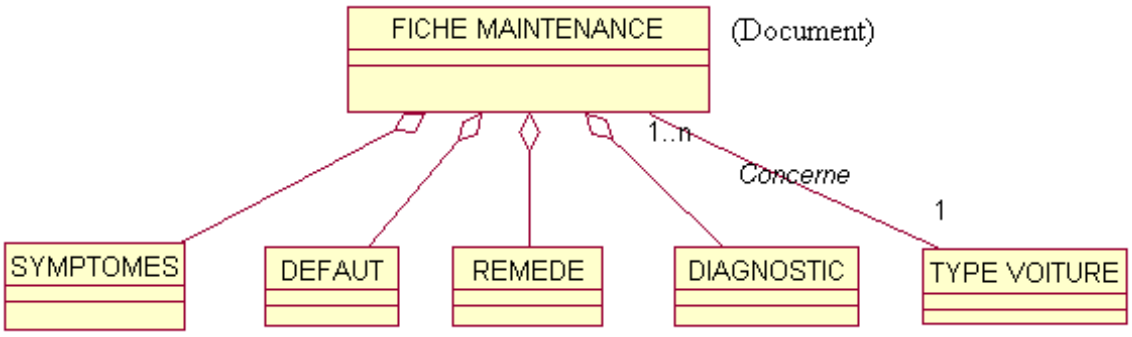

Figure 10 : Structure des documents de maintenance automobile

3.1.3.2. Usage des documents de maintenance automobile

Différents types (concessionnaire, garagiste, mécanicien, conducteur…) et niveaux (débutant, moyen, confirmé, avancé…) d'utilisateurs peuvent vouloir rechercher des documents de maintenance pour un symptôme donné. Ainsi, les fiches de maintenance peuvent être utilisées dans divers contextes dont la maintenance préventive ou corrective, la réparation proprement dite, ou la formation et apprentissage.

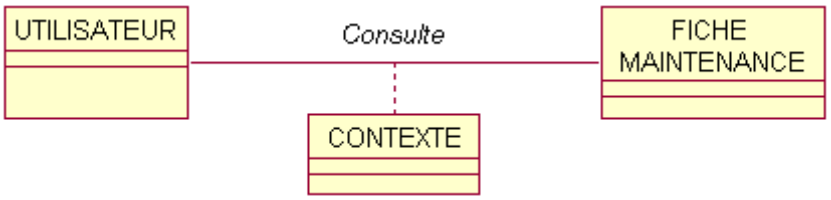

Figure 11 : Usage des documents de maintenance automobile

3.1.3.3. Ontologie du domaine de la maintenance automobile

L'ontologie du domaine de la maintenance automobile regroupe les différentes connaissances nécessaires pour réaliser les tâches de maintenance correctives des pannes automobiles. Les termes décrivant les différents concepts de l'ontologie sont utilisés pour annoter les documents de maintenance automobile présents dans la collection.

Pour ce faire, la partie SYMPTOMES de chaque document est annotée avec les instances de concepts de l'ontologie de maintenance automobile. La structure de cette ontologie est montrée dans la figure 12.

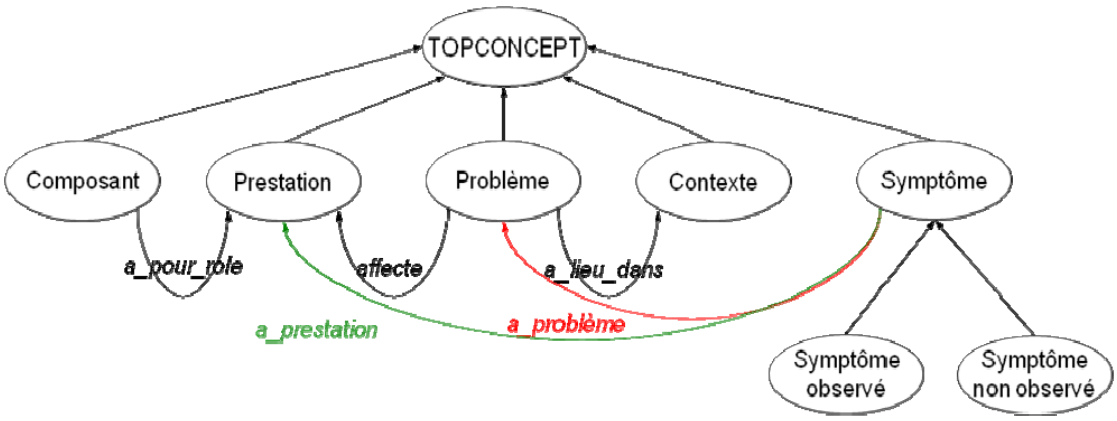

Figure 12 : Structure de l'ontologie du domaine de la maintenance automobile (Reymonet et al., 2009)

Un symptôme donné, qui peut être observé (présence de dysfonctionnement) ou non observé (absence de fonctionnement), concerne un problème quelconque qui a lieu dans un contexte donné. Un contexte décrit un état particulier dans lequel une prestation ne fonctionne pas. Ce même problème affecte une ou plusieurs prestations qui décrivent des fonctionnements de la

voiture. La réalisation d'une prestation est assurée par des composants de la voiture, où un composant est une pièce ou un ensemble de pièces qui participent à la réalisation d'une prestation.

3.1.3.4. Représentation sémantique des contenus des documents de maintenance

Nous trouvons ci-après le modèle de représentation multi-facette des documents de maintenance. Ce modèle est issu de l'instanciation du méta-modèle et regroupe les différentes parties que nous avons présentées dans les sections précédentes.

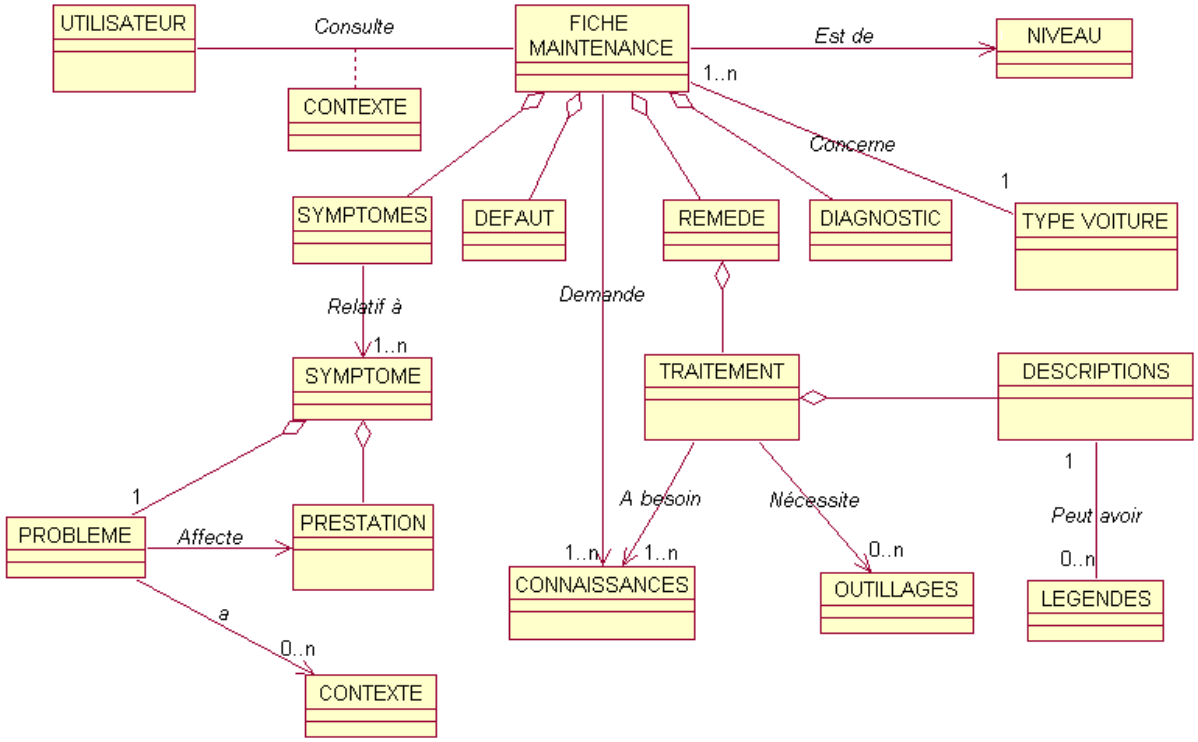

Figure 13 : Modèle complet intégrant les différents aspects de représentation des documents de maintenance automobile.

Toutes les classes constituant le méta-modèle ont été instanciées pour obtenir ce modèle de représentation des documents de maintenance.

La collection de documents est l'ensemble des fiches de maintenance automobile où chaque fiche est composée de plusieurs parties dont le type de voiture, les symptômes, les défauts, le diagnostic, ainsi que le remède.

Chaque remède est composé d'une suite de traitements décrivant l'ensemble des opérations à réaliser sur le type de voiture suivant le diagnostic et le symptôme affiché. Un seul symptôme est traité dans chaque fiche. Chaque symptôme est un problème qui affecte une prestation d'une voiture dans un ou plusieurs contextes.

La classe *TRAITEMENT* regroupe l'ensemble des tâches à effectuer pour traiter les défauts de fonctionnement de la voiture afin de faire disparaître le symptôme. Chaque traitement est composé de plusieurs descriptions qui ne sont que les différentes instructions à suivre pour

chaque étape de la maintenance et qui peuvent être associées avec des légendes qui affichent des photos du composant à traiter.

3.1.4. Modèle de représentation des documents musicaux Braille

Le dernier domaine d'application pour lequel nous avons instancié notre méta-modèle de représentation multi-facette des documents dans le cadre de la RI est la représentation des partitions musicales Braille. Cette application vise à permettre aux malvoyants et non voyants d'accéder aux ressources musicales.

3.1.4.1. Représentation des ressources musicales pour mal et non voyants

La notation musicale Braille est très différente de la notation de musique en noire. Une partition Braille est composée de suites linéaires de symboles représentant des notes, des accords, des parties et des modèles rythmiques. En raison du nombre limité de signes Braille disponibles (64 au total) des éléments musicaux sont produits en utilisant une combinaison d'un ou plusieurs signes Braille. La signification de chaque signe Braille est déterminée suivant son contexte d'utilisation.

Une partition ne se limite pas à une suite de notes. La lecture de la musique prend en compte la dimension verticale. La difficulté du Braille est qu'il faut transcrire un document à "deux dimensions" en une suite linéaire de caractères. En considérant une partition musicale, ce type de lecture n'est pas tout à fait approprié. L'écriture Braille a développé beaucoup de stratégies et quelques symboles spéciaux pour réduire la longueur de texte et le rendre plus facile à lire.

Nous avons défini un langage de codage des partitions musicales Braille dans lequel les concepts Braille spécifiques aux partitions musicales ont été pris en compte (Encelle et al., 2008). Un aperçu de quelques uns de ces concepts est donné dans ce qui suit

Répétition :

Une spécificité majeure de musique Braille par rapport à la musique en noire est l'utilisation vaste de signes de répétition. Cela peut simplifier la lecture, aider à la mémorisation et économiser l'espace. Par exemple, dans le Braille, un caractère spécial (des points 2356) représente une répétition partielle ou complète de mesure.

Séquence :

Pour réduire le temps de lecture en utilisant moins de caractères, une séquence d'éléments similaires en Braille est souvent écrite en doublant l'élément au début de la séquence et en le répétant à la fin. Par exemple, une séquence d'accords de deux notes, qui ont tous le même intervalle de tierce, peut être écrite en doublant le signe d'intervalle après la première note et après la dernière note. Ainsi, pour chaque accord, sauf le premier de la séquence, sa deuxième note (la tierce) n'est pas écrite. Le même système est appliqué pour les séquences de groupe rythmique. Dans le cas d'une succession de plusieurs groupes rythmiques, il est possible de doubler le signe de groupage au début du premier groupe.

Spécification d'octave :

Si l'on veut utiliser une modification d'octave dans une partie de partition, la première note concernée doit être suivie de deux marques d'octaves. La première indique la valeur de

l’octave suivant la position de la note dans la partition en noire, et le second indique la valeur réelle de l’octave.

Durée des notes, accords et silences :

La hauteur d’une note est déterminée par les points 1, 2, 4, et 5 du caractère Braille, et sa durée est spécifiée par la présence ou l’absence des points 3 et/ou 6.

Chaque note ou silence a deux valeurs de durée possibles qui peuvent être déterminées suivant le contexte. Des caractères Braille peuvent préfixer une note pour indiquer précisément sa durée. De plus, pour un accord dont les notes ont la même durée, une note seulement est écrite explicitement. Les autres sont indiquées par leurs intervalles par rapport à cette note.

Copule ou Séparateur de voix (In-accords) :

En Braille, les musiciens peuvent seulement lire horizontalement. Ainsi, les informations verticales doivent être fournies comme une séquence horizontale de caractères. Quand toutes les parties harmoniques ne changent pas en même temps, elles sont montrées en divisant la mesure concernée en voix par l’utilisation des séparateurs de voix, notions qui n’existent pas dans les partitions en noir. Ce symbole indique que les notes qui suivent appartiennent à une autre voix de la même mesure.

Armatures :

Contrairement à la partition en noir, l’armature en Braille indique simplement le nombre de dièses ou de bémols.

Liaisons de phrasé et de prolongation :

Dans les partitions en noir, les liaisons de phrasé et de prolongation sont représentées de la même façon par une ligne qui se trouvent sur ou sous les notes concernées. En Braille, différents caractères sont utilisés suivant le contexte. Parmi ces contextes, nous pouvons citer :

- liaison entre deux notes ou accords,
- liaison de phrasé ayant plus de quatre notes ou accords,
- début et fin de phrasé sur une note,
- début et fin de phrasé de liaison sur une note,
- liaison entre des voix,
- ligne droite entre portées, pour avancement de voix,
- fin de ligne droite,
- liaison ajoutée par l’éditeur dans une partition en noir,
- liaison qui ne se termine pas sur une note,
- liaison sur appogiature courte.

Dispositions :

La différence principale entre la partition en noire et celle en Braille est le concept de dimension spatiale. Pour la partition en noir, les deux dimensions verticale et horizontale sont utilisées pour véhiculer des informations. En Braille, deux dispositions sont disponibles :

- le format section par section dans lequel un groupe de mesures d'une partie d'un instrument alterne avec le même groupe de mesures d'une partie d'un autre instrument,
- le format mesure par mesure dans lequel une mesure d'une partie d'un instrument alterne avec une mesure d'une partie d'un autre instrument.

3.1.4.2. Les codes musicaux Braille

Étant donné que la plupart des codes musicaux existants ne supportent pas (Musicxml, Niff et Midi…) ou supporte en partie les particularités de la musique Braille (Play Code), nous avons été amenés à développer un langage de description des partitions musicales Braille, tout en respectant aussi bien les standards musicaux que les structures spécifiques des partitions Braille. Nous présentons dans la section suivante la structure des partitions Braille.

3.1.4.3. Représentation multi-facette des documents musicaux Braille

Pour pouvoir bénéficier de toutes les particularités des documents musicaux Braille pour leur recherche sémantique, nous avons recensés les différentes facettes que l'on doit considérer pour représenter ce type de document. La facette *Description thématique* représente les différentes connaissances représentant des connaissances musicales d'une part et la musique en générale d'autre part. La description thématique utilise des ontologies musicales[6]. Une autre facette décrit les documents musicaux Braille suivant ses structures et les métadonnées utilisées pour apporter plus de détails dans la partition. La facette Description des représentations statiques et dynamiques présente d'une part le domaine gestuel qui concerne l'interprétation de la partition et d'autre part le graphique qui donne un rendu graphique Braille du contenu de la partition. Enfin, la facette description des usages permet de mettre en relation toutes les autres facettes. En effet, en se basant sur la structure et du contenu musicale de la partition, et suivant la description thématique des contenus et les métadonnées associées, l'interprétation et le rendu graphique de la partition varient en fonction du contexte d'usage de l'utilisateur suivant qu'il soit voyant ou malvoyant, apprenant ou enseignant, et aussi de son niveau en matière de musique en général et de la musique Braille en particulier.

[6] http://musicontology.com/

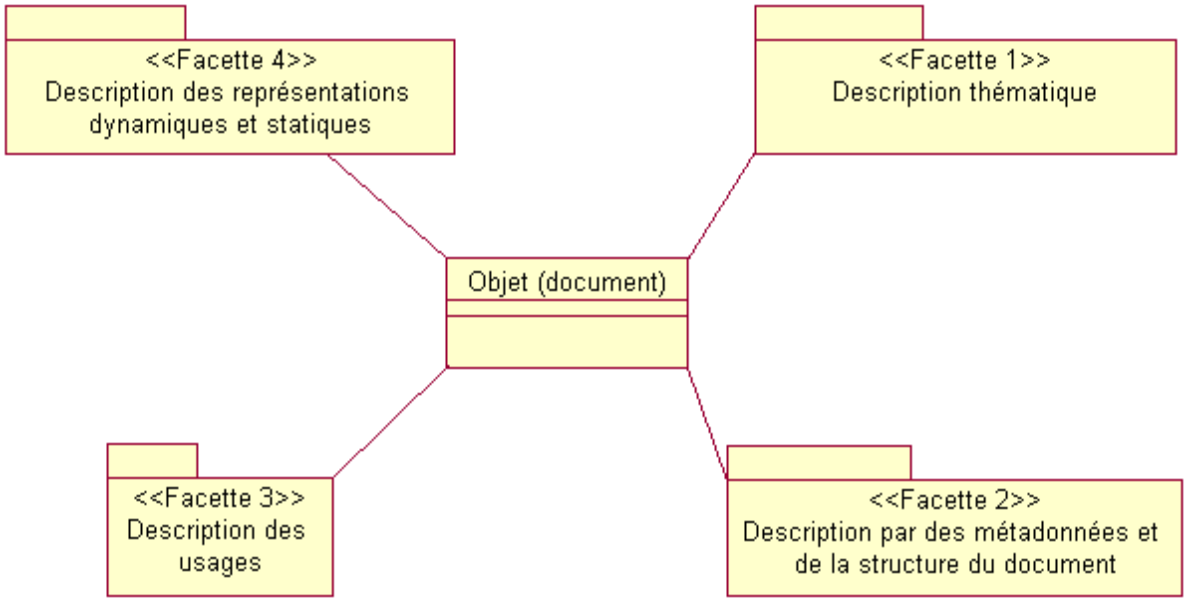

Figure 14 Connaissances utiles pour représenter un document musical Braille et son usage

Partant de cette représentation multi-facette des documents musicaux Braille, nous développons dans la section suivante la structure des partitions Braille.

3.1.4.4. Structure des partitions Braille

Comme indiqué dans la Figure 15, une partition Braille est traitée comme une suite de notes suivies et précédées par plusieurs éléments d'informations comme l'octave, les liaisons, les doigtés et des nuances.....

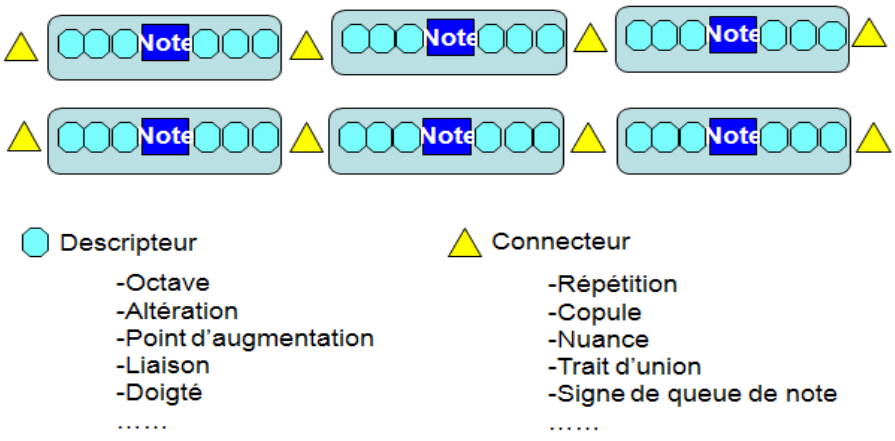

Figure 15 : Structure d'une mesure dans une partition Braille (Encelle et al., 2008).

De la même manière, une partition Braille peut être considérée comme une suite de mesures qui sont connectées au moyen des connecteurs de mesures.

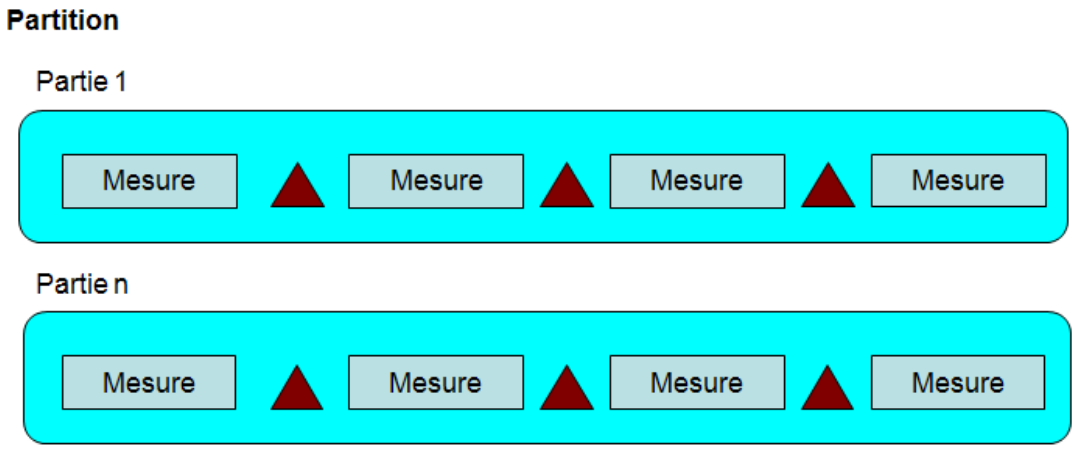

Figure 16 : Structure générale d'une partition Braille.

3.1.4.5. Contexte et description des représentations dynamiques et statiques

La description des représentations dynamiques correspond à l'interprétation MIDI de la partition musicale Braille donnée tandis que la représentation statique concerne la représentation graphique et Braille de la partition concernée.

Pour une partition Braille donnée, plusieurs représentations sont possibles suivant le contexte d'utilisation de l'utilisateur. La même partition est présentée différemment selon que l'utilisateur soit enseignant, apprenant, voyant ou malvoyant. De même, selon le niveau musical de l'utilisateur (débutant, moyen, avancé), l'interprétation dynamique (exécution) du même morceau peut varier.

3.1.4.6. Modèle de représentation multi-facette des partitions musicales Braille.

À chaque partition Braille sont associées des métadonnées qui permettent de décrire les informations concernant la partition (par exemple : Titre, Genre, Auteur, Compositeur, Transcripteur, Date d'édition).

Des métadonnées et annotations sont aussi associées à chaque partie et sections d'une partition afin de permettre aux utilisateurs d'apporter des informations supplémentaires comme des annotations personnelles.

Une partition Braille est composée d'une ou plusieurs parties (chaque partie correspond à un instrument donné). Chaque partie est composée de sections (elles-mêmes composées d'ensemble de mesures), de lyriques et de symboles d'accord.

Chaque mesure est reliée avec d'autres mesures à l'aide de connecteurs. A l'intérieur de chaque mesure se trouvent les éléments musicaux dont les notes, les silences, les liaisons et les séparateurs de voix (Copule). Chacun de ces éléments musicaux peut être décrit à l'aide des pré-descripteurs et post-descripteurs qui précisent les fonctions de chaque élément dans le contexte d'utilisation car un même symbole Braille peut avoir plusieurs significations.

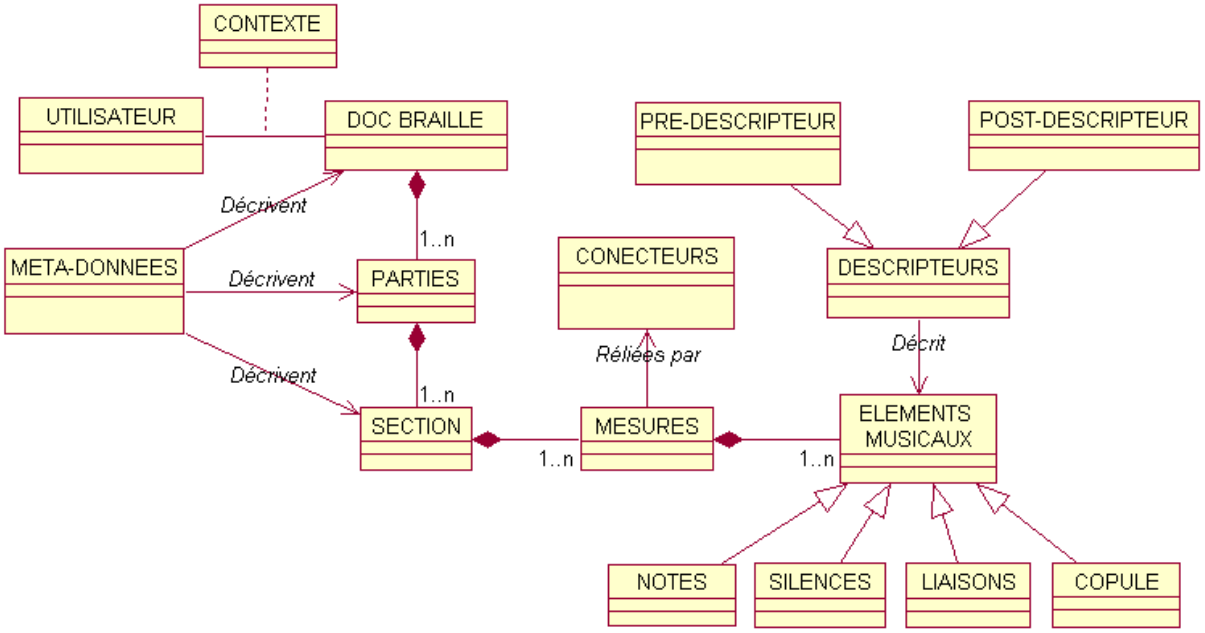

Figure 17 : Modèle de représentation multi-facette des partitions musicales Braille

À partir de ce modèle de représentation des partitions Braille, nous avons défini le schéma XML du langage correspondant BMML (Braille Music Markup Language) qui est détaillé dans la section 4.3.3.2.

3.1.5. Conclusion

Afin de parvenir à réaliser des recherches sémantiques sur les documents numériques, il est indispensable de pouvoir faire ressortir les différentes sémantiques véhiculées dans ces documents. Étant donné que la sémantique des documents ne dépend pas seulement des textes contenus dans ce document mais également de la structure, du format et du média utilisé, nous avons proposé un méta-modèle qui représente les différentes facettes des documents. Ce méta-modèle a été instancié dans différents domaines d'application : l'apprentissage en ligne, la maintenance automobile et la musique Braille afin d'obtenir le modèle de représentation multi-facette respectivement des objets pédagogiques, des documents de maintenance et des partitions musicales Braille.

Le modèle de chacun de ces domaines est une instanciation de toute ou partie du méta-modèle. Les modèles de représentation multi-facette des documents regroupent la composition et la structure des documents, leurs descriptions par des facettes et des concepts de l'ontologie du domaine, ainsi que les tâches associées aux documents. De la sorte, ces modèles permettent de garantir la cohérence entre documents, leurs représentations et les vocabulaires d'indexation.

Dans la section suivante, nous traitons l'indexation sémantique des documents tirant profit des différentes facettes des documents, de leurs structurations et des concepts de l'ontologie de domaine.

3.2. Indexation sémantique de documents

L'indexation sémantique des documents consiste à annoter les contenus de tout ou une partie des documents par des concepts d'une ontologie de référence en vue de leur recherche sémantique ultérieure. L'annotation sémantique avec des concepts d'une ontologie permet de lever l'incertitude sur les termes ambigus trouvés dans le document. Des relations (de subsomption ou de type lien sémantique) peuvent exister entre les différents concepts annotant le même document. Une annotation donnée peut se faire donc soit avec un concept indépendant, soit à l'aide d'un triplet regroupant deux concepts et une relation.

Dans cette section, nous allons présenter notre étude sur la mise en œuvre d'un SRI sémantique qui nous a permis d'une part de recenser les différentes contraintes et différents paramètres à prendre en compte et d'autre part, à partir de ces derniers, de proposer un modèle d'indexation dynamique à base d'ontologie (Hubert et al., 2009). Notre modèle peut être appliqué dans les différentes approches d'indexation sémantique dont l'indexation à partir du texte et celle à partir d'ontologie.

Cette section présente les contraintes de l'indexation (3.2.1), le modèle d'indexation sémantique à base d'ontologies (3.2.2), l'annotation de documents par des graphes de concepts (3.2.3), l'indexation à partir du texte de document (3.2.4), l'indexation à partir d'ontologies (3.2.5) et la pondération des concepts annotant les documents (3.2.6).

3.2.1. Contraintes de l'indexation

Un SRI est jugé au travers de ses performances en termes de satisfaction des utilisateurs sur la pertinence des documents retrouvés, le temps de réponse aux requêtes utilisateurs et la disponibilité du système. Les performances des SRI dépendent des algorithmes d'indexation utilisés et surtout de la structure des listes inversées qui associent les documents avec les termes d'indexation. Ainsi, afin de définir les structures de données à utiliser dans les listes inversées, plusieurs paramètres qui entrent en jeu dans la performance des SRI doivent être pris en compte. Parmi ces paramètres, nous pouvons citer : la taille du corpus (nombre de documents constituant la collection), la fréquence de mise à jour de la collection et le format des documents.

La mise à jour de la collection demande la ré-indexation des documents. La taille du corpus affecte donc la durée de ré-indexation, et qui peut engendrer une lenteur du système, voire son indisponibilité pendant un certain temps. De même, la fréquence de mise à jour de la collection a un impact sur la disponibilité de l'index. En effet, plus la fréquence de mise à jour de la collection est élevée, moins l'index est disponible pour la RI car il est à tout moment en cours de modification. Enfin, le format des documents affecte le temps d'indexation car les durées d'extraction des termes d'un document ne sont pas les mêmes pour tous les formats.

Ces paramètres liés à la collection se combinent avec les exigences des utilisateurs qui souhaitent obtenir des documents pertinents dans un meilleur délai et les principes utilisés en RI au moment de l'indexation. Ainsi, notre modèle prend en compte les contraintes dont la granularité de l'indexation, la disponibilité de la collection et de l'index, l'évolution de la collection, le temps de réponse, la pondération des termes d'indexation et la recherche sémantique. Ces différentes contraintes sont présentées dans les sous-sections suivantes.

3.2.1.1. Granularité de l'indexation

L'unité d'indexation et de restitution est soit un document complet soit une ou des parties de documents bien délimitée. Cependant, les positions des occurrences des termes représentant des concepts dans un document doivent être mémorisées de façon à permettre d'afficher à l'utilisateur un aperçu du document au moment de la visualisation des résultats d'une recherche.

3.2.1.2. Disponibilité de la collection et de l'index

La minimisation du délai de mise à jour dynamique d'index permet d'améliorer la disponibilité de la collection et de l'index. Dans les SRI actuels, suivant la taille du corpus, la mise à jour de l'index peut prendre des heures de traitement. Ceci limite donc les ajouts de nouveaux documents ou les modifications de documents dans le corpus et leur prise en compte dans les index.

3.2.1.3. Évolution de la collection

Selon notre approche, lorsqu'un document est ajouté, seules les données statistiques du nouveau document et de ses expressions (fréquences d'apparition suivant le type de texte) sont mises à jour. Le calcul des poids des expressions et des concepts se fait au moment de l'évaluation de la requête. De plus, la mise à jour des données suite à une modification de la collection ne s'effectue qu'au niveau des documents concernés.

3.2.1.4. Temps de réponse

La minimisation du temps de réponse lors du traitement des requêtes est une des contraintes à respecter par tout SRI. Le temps qui s'écoule entre la saisie de la requête et l'affichage des résultats doit se situer dans la limite acceptable par l'utilisateur. Le fait de traiter dynamiquement l'indexation des documents ne doit pas pénaliser le temps de réponse du système lors d'une recherche de documents. La minimisation du temps d'indexation assure la disponibilité permanente de l'index. A son tour, la disponibilité de l'index permet d'évaluer les requêtes utilisateurs à tout moment. Dans notre modèle, toutes les informations statistiques et sémantiques qui servent à l'évaluation des requêtes sont accessibles dans une base de données des index.

3.2.1.5. Pondération des termes d'indexation

La pondération des termes d'indexation permet de se rendre compte de leur pouvoir discriminant. La pondération des expressions dans les index tient compte de la mesure tf*idf. Ce poids associé aux termes d'indexation (Robertson, 1976) est utilisé lors de l'étape d'appariement d'une requête avec les documents. Tf (Term Frequency) est la fréquence d'apparition d'un terme dans le document et Idf (Inverse Document Frequency) est la valeur de l'importance du terme dans l'ensemble de la collection. Dans notre modèle, la pondération tient également compte de certains éléments associés aux expressions dans le document comme leur mise en forme.

3.2.1.6. Recherche sémantique

La recherche sémantique des documents se fait sur leur contenu. Le système recherche dans l'ontologie les concepts qui correspondent aux termes de la requête puis restitue les documents qui sont indexés par ces concepts.

Les structures de données que nous allons présenter ci-après ont été conçues pour répondre à ces contraintes et objectifs. Elles contiendront donc les éléments de l'index et aussi de

l'ontologie qui sert de référence à l'indexation. Notons que le format des documents de la collection n'affecte pas les structures de données.

3.2.2. Modèle d'indexation sémantique à base d'ontologies

Nous avons défini un modèle de données servant de socle à la mise en place d'un SRI basé sur une indexation sémantique par ontologie. Ce modèle présenté dans la *Figure 18* prend notamment en compte le double objectif d'actualisation dynamique des listes inversées et d'utilisation d'ontologies lors de l'indexation.

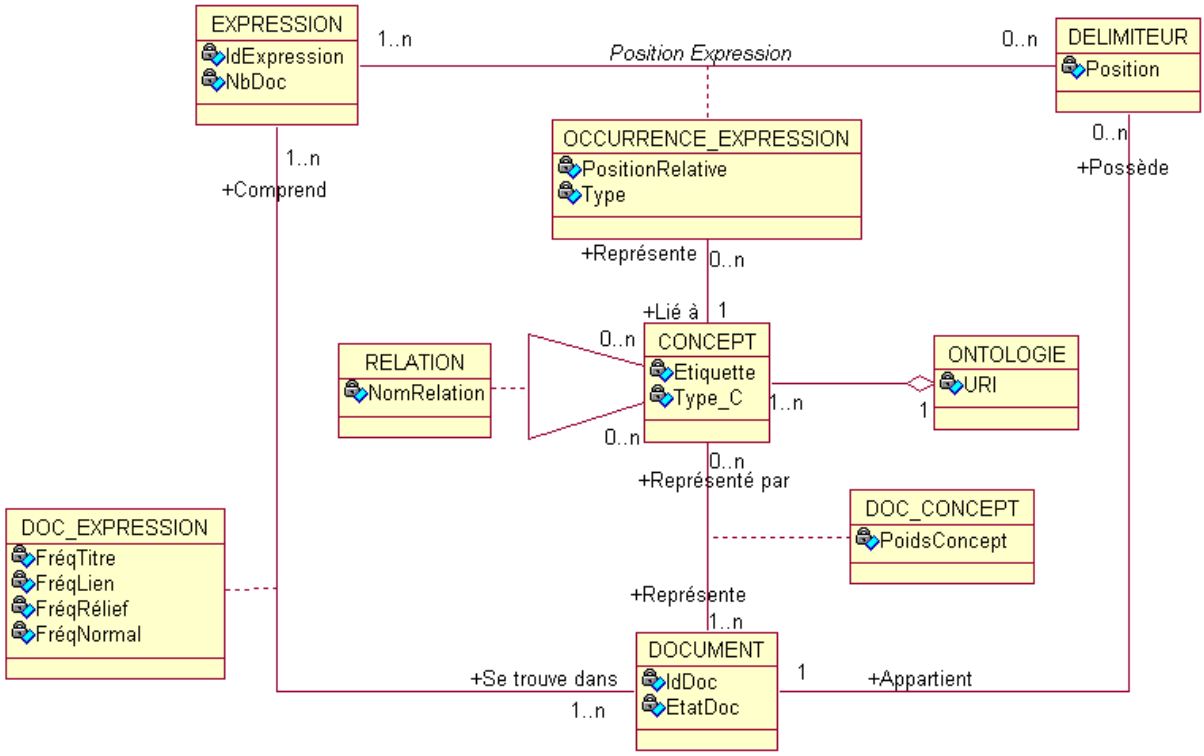

Figure 18 : Diagramme de classes représentant les données utilisées pour l'indexation. (Hubert et al., 2009)

La classe *DOCUMENT* représente les unités d'indexation. A chaque document est associé son identifiant (*IdDoc d'un document correspondant à son URI - Uniform Resource Identifier*), unique référence dans l'ensemble du système. Trois états possibles, impliquant des traitements différents, sont distingués pour un document. L'état d'un document peut être *normal* (cas des documents intégrés dans le système avec indexation à jour), *mis à jour* (cas des documents dont le contenu a été modifié depuis la précédente indexation) ou bien *effacé* (le document est retiré logiquement du système depuis la précédente indexation).

Chaque document est subdivisé en plusieurs sections séparées par des marqueurs logiques. La classe *DELIMITEUR* représente ces marqueurs en précisant leur *Position* absolue dans le document.

Par ailleurs, chaque document est considéré comme un ensemble d'expressions décrites au travers de la classe *EXPRESSION*. Une expression est un groupe de termes représentant un concept du domaine. Le nombre de documents (*NbDoc*) où apparaît chaque expression est

conservée. Chaque OCCURRENCE D'UNE EXPRESSION apparaissant dans un document est repérée par sa position relative (*PositionRelative*) par rapport à un délimiteur du document. Ceci permet d'associer une annotation ancrée dans une partie du texte du document. .

La propriété *Type* conserve l'importance du texte incluant l'occurrence de l'expression (c'est-à-dire si celle-ci apparaît dans un titre, dans un lien hypertexte, dans un texte mis en relief ou dans un texte normal). La mémorisation des positions des occurrences des expressions (*PositionRelative*) dans les documents offre des possibilités de localisation précise des index pour l'utilisateur.

De plus, chaque occurrence d'une expression peut être associée à un concept d'une ontologie de référence. La classe CONCEPT décrit les instances des concepts identifiées par leur *Etiquette* telle que définie dans l'ontologie. Cela permet de faire le lien avec la définition de l'ontologie repérée par son *URI* (classe *ONTOLOGIE*).

Une instance de concept peut être liée à d'autres par des instances de la classe RELATION (ces relations étant définies entre concepts dans l'ontologie). Plusieurs relations pouvant exister entre deux concepts, il est nécessaire de conserver le nom de la relation concernée (Est-un, etc) dans la propriété *NomRelation*.

La classe d'associations *DOC_EXPRESSION* rassemble les données statistiques concernant les expressions et leur apparition dans les documents (fréquence d'apparition dans les titres, dans des liens, dans des textes en relief, dans des textes normaux).

Enfin, la classe d'association *DOC_CONCEPT* précise les poids des concepts associés à un document. Un concept n'est pas forcément associé à une expression d'un document, il peut être affecté explicitement au document sans posséder un ancrage textuel.

Notre modèle de données permet l'indexation des documents, soit par des groupes de concepts (graphes de concepts) qui sont reliés entre eux, soit par des concepts isolés.

3.2.3. Annotation par des graphes de concepts ou des concepts isolés

L'annotation d'un document peut se faire par rapport à un ancrage dans une partie du texte du document soit sur le document entier, ou bien les deux en même temps. Le premier type d'annotation permet ultérieurement de réaliser des recherches d'occurrences de concepts dans le document tandis que le second ne permet que de retrouver, parmi les documents présents dans le corpus, ceux qui ont été annotés par les concepts de la requête sans préciser ou sans avoir trouvé dans les textes des termes associés aux concepts.

3.2.3.1. Annotation de parties de texte

Une partie de texte dans un document peut être annotée soit par des concepts indépendants ou par des graphes de concepts. Chaque concept est alors associé à une occurrence de texte dans la partie courante du document. De même, un document peut être annoté sur une ou plusieurs parties de textes de son contenu.

3.2.3.2. Annotation d'un document via des métadonnées

Des concepts peuvent annoter un document sans être associés avec une occurrence de texte dans le document. Dans ce cas, les concepts sont considérés comme des métadonnées associées au document et ces concepts sont stockés dans la classe d'objet *DOC_CONCEPT*.

Moyennant notre modèle d'indexation dynamique à base d'ontologies, nous pouvons indexer un document du corpus de deux manières ; soit l'indexation se fait en partant du texte du

document vers l'ontologie, soit de l'ontologie vers le texte du document. Les détails de ces deux approches sont exposés dans les sections suivantes.

3.2.4. Indexation à partir du texte

L'indexation à partir du texte est un processus d'indexation sémantique qui consiste à associer un graphe de concepts ou un ensemble de concepts isolés à un texte. Il s'agit d'abord de trouver dans l'ontologie les concepts candidats qui pourraient être désignés par le terme présent dans le document. Sachant qu'un terme donné peut représenter divers concepts dans l'ontologie, une étape de désambigüisation peut être nécessaire. Dans le cas d'une indexation par graphe de concepts, l'extraction de la relation correspondante sera également nécessaire. Ces deux phases sont décrites dans les sous sections suivantes.

3.2.4.1. Recherche de concepts de l'ontologie

L'objectif est de trouver, à partir du terme présent dans le document, le concept correspondant dans l'ontologie. Dans ce cas, le terme est comparé aux termes qui dénotent les concepts dans l'ontologie afin de trouver des concepts candidats. Il reste à trouver le bon concept parmi ces différents candidats afin de lever l'ambigüité potentielle. Pour ce faire, la connaissance des autres concepts candidats présents dans le texte, moyennant les relations entre concepts dans l'ontologie, permettra d'affiner la sélection des bons concepts. La présence d'une instance de relation dans le document permet de ne retenir que les concepts associés avec la relation dans l'ontologie.

3.2.4.2. Recherche de relations entre concepts

Des termes peuvent désigner des noms de relations dans l'ontologie. Ainsi, afin de spécifier lesquels des concepts candidats présélectionnés sur des termes pour un bout de texte donné doivent être retenus, nous devons non seulement trouver dans l'ontologie la relation correspondant au terme du document mais également associer la relation avec ce terme. La relation une fois identifiée permet de cerner les deux concepts qu'elle relie. Nous n'avons plus qu'à retenir parmi les concepts candidats associés aux termes, du document, ceux qui correspondent aux concepts reliés par la relation.

3.2.5. Indexation à partir de l'ontologie

L'indexation à partir de l'ontologie est un autre moyen d'indexation sémantique qui vise à rechercher dans le document courant les occurrences des instances de concepts et de relations en se basant sur un bout d'ontologie. Ce type d'indexation est donc fait en deux étapes dont la recherche des occurrences de concepts suivie de celle des occurrences des relations.

3.2.5.1. Recherche d'occurrences de concepts dans le texte

Pour un concept donné dans l'ontologie, nous recherchons dans chaque partie du document les occurrences des termes qui désignent ce concept. Comme un concept peut être désigné par plusieurs termes, qui peuvent désigner à leur tour plusieurs concepts, des termes polysèmes peuvent être un candidat à associer au concept courant pour un bout de texte donné dans le document. De ce fait, pour déterminer lesquels de ces termes candidats correspondent au concept courant, l'analyse des occurrences des relations reliant ce concept avec les

occurrences d'autres concepts est utilisé. Les termes ainsi déterminés seront utilisés pour l'annotation.

3.2.5.2. Recherche d'occurrences de relations dans le texte

Pour les concepts de l'ontologie qui sont en relation avec d'autres concepts, la recherche des instances de la relation reliant ces concepts permet de filtrer les concepts à associer au terme, et ainsi de désambigüiser les termes polysèmes parmi les candidats précédemment sélectionnés. Ainsi, pour un bout de texte donné, si plusieurs concepts candidats sont présents pour un terme donné, sont gardés ceux qui dénotent les concepts reliés par la relation courante. Le terme associé avec le concept retenu est en relation avec un autre terme associé avec un autre concept via l'instance de la relation. Les deux concepts sont en relation dans l'ontologie, par la relation qui subsume l'instance de relation trouvé dans le texte.

L'association des concepts et relations avec des termes dans un document permet de rechercher sémantiquement ces documents au cours d'une session de RI. Cependant cela ne suffit pas pour pouvoir classer les documents potentiellement pertinents retrouvés. De manière à permettre ce classement des documents résultats, nous proposons ci-après de pondérer les concepts associés à chaque document.

3.2.6. Pondération des concepts

La pondération des concepts consiste à donner plus d'importance à certains concepts par rapport à d'autres par association de poids à chaque concept. Le calcul des poids des concepts se base sur le principe Tf*Idf (Cf section 2.2.1.1, page 15). De plus, la pondération tient compte de certains éléments associés aux expressions dans le document comme leur propriété de mise en forme. En effet, les concepts présents dans un titre sont considérés comme plus importants que ceux dans les corps de textes. De même, les concepts présents dans un texte mis en relief (gras, souligné, etc…) sont considérés comme plus importants que ceux dans un texte sans mise en forme. Notre hypothèse est de calculer le poids de chaque concept en fonction des rôles joués par chaque occurrence des termes associés au concept.

Le poids final d'un concept est la moyenne pondérée des poids de chaque type d'occurrence (titre, gras, souligné, normal, lien…) des termes associés au concept.

$$Poids_{Concept}(C) = \frac{\sum Coeff(T) \times Poids_{Type}(T)}{\sum Coeff(T)} \qquad \textbf{(21)}$$

$$\text{Où } Poids_{Type}(t) = Tf(C,T) \times Idf(C,T) \qquad \textbf{(22)}$$

Avec $Tf(C,T)$: Nombre d'occurrences des termes associés au concept C en ayant le type T de mise en forme.

$Idf(C,T)$: La fréquence absolue des termes associés au concept C ayant une mise en forme de type T.

$Coeff(T)$: Coefficient associé au type de mise en forme T de façon à pondérer l'importance de chaque type de mise en forme.

Dans un environnement où les documents du corpus ainsi que les connaissances inhérentes au domaine évoluent fréquemment dans le temps, nous proposons de ne calculer les poids des concepts qu'au moment de l'appariement des requêtes et documents (Hubert et al., 2009) afin de garantir la disponibilité des index et de permettre ainsi l'accès aux documents. Nous détaillons dans la section 3.3 les dynamiques de la RI sémantique.

3.2.7. Conclusion

Nous avons proposé des stratégies permettant l'association des termes du document avec les concepts de l'ontologie de deux manières : soit en partant des termes du document et en cherchant dans l'ontologie les concepts et relations correspondants, soit en partant des concepts et relations de l'ontologie et en cherchant dans les documents leurs occurrences respectives.

Notre modèle d'indexation a été conçu pour permettre d'annoter aussi bien les parties de documents que les documents entiers. Ceci facilite aussi la maintenance des index en cas de modification des parties de documents. Ainsi, la granularité de recherche et de restitution de document peut se faire à deux niveaux : soient des parties de documents soient des documents entiers.

Afin d'accorder une importance relative aux concepts par rapport au document, nous avons choisi pour la pondération des concepts indexant le document. Notre méthode de pondération de concept repose non seulement sur le principe de Tf-Idf mais également sur les propriétés de mise en forme des textes dans les documents. Ainsi, Tf-Idf est appliqué pour chaque type de mise en forme afin de donner plus d'importance aux documents comportant les termes dans les titres par rapport à ceux comportant les termes dans les corps de documents.

L'une des problématiques de la RI sémantique est l'évolution dans le temps d'une part des documents de la collection et d'autre part de l'ontologie de domaine qui sert de base de référence des concepts utilisés pendant la phase d'indexation. Ces évolutions nécessitent la remise en cause des index courants pour garder la cohérence entre index et documents. Nous traitons dans la section 3.3 la dynamique de l'indexation sémantique tant au niveau global de la collection qu'au niveau de chaque document de la collection.

3.3. Dynamique de l'indexation sémantique des contenus

Quelque soit la facette considérée, la représentation du document est sujette à modification. Dans cette section, nous considérons la facette « description thématique », en notant que les propositions peuvent être généralisées à l'ensemble des facettes.

L'indexation sémantique des contenus est sujette à trois types de dynamiques différents : la dynamique de la collection, la dynamique du document et la dynamique de l'ontologie de référence. Nous nous focalisons nos études sur la dynamique de la collection et celle du document. Dans nos travaux, nous nous inspirons des différentes techniques qui visent à indexer les documents sur le web et les adaptons à des collections indexées par des ontologies. Ainsi, la généralisation des méthodes proposées par Google (Page et Brin, 1998) associée avec la méthode *Délimiteur-Diff* (Lim et al., 2007) nous permettent de gérer dynamiquement l'évolution de l'indexation des documents d'une collection.

Nous détaillons dans les sections suivantes nos solutions pour traiter la dynamique de la collection (3.3.1) ainsi que celle des documents (3.3.2).

3.3.1. Dynamique de la collection

La dynamique de la collection consiste en l'ajout de nouveaux documents ou la suppression de documents du corpus. Ces événements nécessitent la mise à jour des index car l'ajout de nouveaux documents peut apporter de nouveaux concepts tandis que la suppression de documents peut engendrer soit la diminution de sa fréquence d'apparition dans le corpus voire sa disparition du corpus. De plus, les nouveaux documents doivent être indexés pour pouvoir être retrouvés. En outre, ces suppressions et ajouts peuvent avoir des impacts sur l'ontologie de référence. Nous ne nous intéressons pas à ce dernier cas.

3.3.1.1. Ajout de nouveaux documents

A l'arrivée d'un nouveau document, la mise à jour de l'index et des structures associées (cf. Figure 18) est réalisée selon l'algorithme 1 :

```
Entrées : Un nouveau document
Sorties : Index mis à jour, Document ajouté dans le corpus
Début
        Créer une instance de DOCUMENT à l'état (EtatDoc) normal.
        Délimiter le nouveau document en blocs de paragraphes
        Pour chaque bloc lié à un délimiteur Faire   /* Intégration des nouveaux blocs  */
                Créer une instance dans DELIMITEUR
                Extraire les expressions décrivant le bloc
                Pour chaque expression Faire  /* Ajout d'expression */
                        Si l'expression n'existe pas dans la classe EXPRESSION Alors
                                Créer une nouvelle instance d'expression
                                Créer une nouvelle instance de DOC_EXPRESSION
                        FinSi
                        Si instance de DOC_EXPRESSION n'existe pas Alors
                                Créer une nouvelle instance de DOC_EXPRESSION liée à
                                l'expression.
                                Mettre à jour la valeur de NbDoc dans EXPRESSION
                        Sinon
                                Mettre à jour les propriétés (fréquences d'apparition) pour l'instance
                                de DOC_EXPRESSION
                        FinSi
```

Créer une instance de la classe d'association OCCURRENCE_EXPRESSION liée aux instances d'expression et de délimiteur en cours.
Identifier l'instance de CONCEPT correspondant à l'occurrence d'expression et les éventuelles instances de RELATION auxquelles l'instance de CONCEPT participe
FinPour
Identifier les éventuelles instances de RELATION reliant les instances de CONCEPT
FinPour
Fin

Algorithme 1 : Prise en compte de l'ajout d'un nouveau document

3.3.1.2. Suppression de documents

Pour un document supprimé, son statut est changé en « Effacé » avant de mettre à jour les différentes informations relatives au document dans la base. Ainsi, ce document ne sera plus pris en compte par les requêtes. Après les mises à jour des données du document, le document pourra être supprimé physiquement de la collection. L'algorithme 2 correspond à la suppression d'un document.

Entrées : Un document à supprimer, Un corpus
Sorties : Index mis à jour, Document retiré du corpus
Début
Changer l'état du *DOCUMENT* à l'état (*EtatDoc*) Effacé.
/* Le document sera ignoré par toutes les requêtes */
Pour chaque délimiteur du document dans *DELIMITEUR* Faire
Supprimer toutes les occurrences de *OCCURRENCE_EXPRESSION* liées au délimiteur
Supprimer le délimiteur
FinPour
Pour chaque instance de *DOC_EXPRESSION* liée au document Faire
Décrémenter *NbDoc* de l'expression correspondante dans *EXPRESSION*
Si *NbDoc* = 0 Alors Supprimer l'expression dans *EXPRESSION*
Supprimer l'occurrence de *DOC_EXPRESSION*
FinPour
Supprimer les instances de RELATION reliant les instances de CONCEPT du document
Supprimer les occurrences de *DOC_CONCEPT* relatif au document
Supprimer les instances de CONCEPT du document
Supprimer le document dans *DOCUMENT*
Supprimer physiquement le document de la collection
Fin

Algorithme 2 : Suppression d'un document.

3.3.2. Dynamique d'un document

La dynamique d'un document concerne l'évolution de ses contenus dans le temps. Des ajouts, modifications ou suppressions de textes peuvent être faits sur un document donné. Ces événements nécessitent la mise à jour des index, afin de garder la cohérence entre l'index et

les documents du corpus, car ils peuvent être la source d'apport de nouveaux concepts, de suppression de concepts ou de modification des fréquences d'apparition de concepts.

Pour la modification d'un document, son statut passe à l'état « Modifié » avant la phase de mise à jour des données relatives aux expressions du document. Les éventuelles requêtes faisant appel au document en cours de modification utilisent les anciennes valeurs dans la base en attendant la fin des mises à jour des données relatives au document modifié.

L'état du document redevient « Normal » après les mises à jour des données.

3.3.2.1. Ajout de texte

L'ajout de nouveaux textes dans un document peut engendrer les événements suivants :

- annotation par d'autres concepts de l'ontologie,
- annotation par un nouveau concept (qui n'était pas dans l'ontologie) dans le document,
- annotation par une nouvelle relation reliant des concepts.

L'ajout de textes contenant des expressions qui n'étaient pas déjà présentes dans le document déjà indexé a pour conséquence de modifier l'indexation comme le montre l'algorithme 3.

```
Entrées : Document modifié
Sorties : Index mis à jour
Début
        Changer l'état du DOCUMENT à l'état Modifié.
        /* L'affectation du nouveau texte au bloc dans lequel il est inséré */
        Pour chaque instance d'EXPRESSION dans le nouveau texte faire
                Créer une occurrence dans OCCURRENCE_EXPRESSION
                Si l'instance de DOC_EXPRESSION n'existe pas Alors
                        Créer une nouvelle instance de DOC_EXPRESSION liée à l'expression
                        Incrémenter la valeur de NbDoc dans EXPRESSION
                FinSi
                Mettre à jour les propriétés (fréquences d'apparition) pour l'instance de
                DOC_EXPRESSION
                Mettre à jour les positions des délimiteurs (du document) dont la valeur de la propriété
                position est supérieure à celle du délimiteur du bloc contenant la nouvelle expression.

                Si la taille du bloc devient trop importante (le double de la taille normale d'un bloc)
                Alors
                        /* le bloc courant est éclaté en deux. */
                        Créer une nouvelle occurrence de délimiteur dans DELIMITEUR
                        Pour chaque expression de la deuxième partie du bloc faire
                                Affecter l'occurrence d'expression au nouveau délimiteur
                                Mettre à jour la position relative par rapport au nouveau délimiteur
                        FinPour
                FinSi
        FinPour
        Changer l'état du DOCUMENT à l'état Normal
Fin
```

Algorithme 3 : Modification d'un document – ajout d'expressions

Suite à des successions d'ajouts d'expression, la taille du bloc peut devenir trop importante. Cela peut affecter la durée de mise à jour des instances de *DOC_EXPRESSION* dans le cas où la nouvelle expression est insérée au début du bloc. L'éclatement d'un grand bloc permet de réduire le nombre de mises à jour à effectuer dans *DOC_EXPRESSION*.

3.3.2.2. Suppression de texte

La suppression de texte dans un document peut générer les événements suivants :

- suppression des instances de relation ayant relié les concepts annotant la partie de texte,
- suppression de concepts annotant la partie de texte,

La suppression d'une expression entraîne la modification des instances du modèle comme l'indique l'algorithme 4.

```
Entrées : Document modifié
Sorties : Index mis à jour
Début
        Changer l'état du DOCUMENT à l'état Modifié
        Supprimer l'occurrence dans OCCURRENCE_EXPRESSION
        Décrémenter la propriété fréquence correspondante dans DOC_EXPRESSION
        Si la somme des fréquences est égale à zéro alors
                Supprimer l'entrée dans DOC_EXPRESSION
                Décrémenter NbDoc dans EXPRESSION
                Si Nbdoc=0 alors supprimer l'expression dans EXPRESSION
        FinSi

        Si la Taille du bloc courant < Seuil et le nombre de blocs du document >=2 alors
                /* Bloc devenu trop petit suite à plusieurs suppressions*/
                Si le bloc courant correspond au premier délimiteur du document alors Prendre le bloc
                suivant comme bloc courant
                Pour chaque expression du bloc courant Faire
                        Affecter l'expression au bloc précédent
                        Mettre à jour sa position relative par rapport au bloc précédent
                FinPour
                Supprimer le délimiteur du bloc dans DELIMITEUR
        FinSi
        Changer l'état du DOCUMENT à l'état Normal
Fin
```

Algorithme 4 : Modification d'un document – suppression d'expressions

Pour limiter le nombre de blocs dans un document, les blocs de petite taille seront fusionnés avec un autre bloc car un nombre important de blocs entraîne plusieurs mises à jour dans *DELIMITEUR* à chaque ajout d'un nouveau bloc.

L'affectation des expressions à la fin du bloc précédent permet d'éviter de modifier toutes les positions relatives des expressions du bloc suivant.

3.3.2.3. Modification d'un bloc de texte

La modification d'un bloc de texte se traduit par la séquence :

- suppression d'expressions,
- ajout d'expressions.

Ceci implique les opérations suivantes :

- suppression de concepts,
- suppression de relations,
- annotation par un autre concept de l'ontologie,
- annotation du document par un nouveau concept non présent dans l'ontologie. Ce cas nécessite la mise à jour de l'ontologie de référence,
- éventuellement, annotation du document par une relation reliant deux concepts.

Dans tous les cas (Ajout, Suppression ou Mise à jour de documents de la collection), la mise à jour dynamique de l'index n'est réalisée que sur les documents concernés. Cela permet de diminuer le temps d'indexation, et ainsi d'augmenter la disponibilité de la collection à tout moment. De plus, l'indexation dynamique permet de conserver une cohérence entre collection et index, ce qui permet au système de trouver les documents pertinents à tout moment.

3.3.3. Conclusion

L'indexation d'une collection de documents peut prendre des heures si le nombre de documents à indexer est très important et/ou les documents de la collection sont de très grandes tailles. A cela s'ajoute la dynamique de l'indexation de documents qui consiste à la mise à jour des index à chaque arrivée, modification ou suppression de documents de la collection. Ceci se produit dans le cadre d'indexation des documents sur web ou celle des documents d'entreprises dont les contenus varient fréquemment dans le temps.

Pendant la période d'indexation, les documents ne sont pas accessibles aux utilisateurs car les index qui permettent de les chercher sont en cours de modification. Notre modèle d'indexation sémantique à base d'ontologies permet non seulement l'indexation sémantique des contenus des documents, mais résout également la problématique de non accessibilité des documents pendant les phases d'indexation. Ainsi, même si des modifications s'appliquent fréquemment aux contenus des documents, la collection reste accessible.

Quant à l'ontologie de domaine qui sert de référence à l'indexation des documents, en tant que représentation formelle des connaissances, elle peut évoluer dans le temps. Cette mise à jour de l'ontologie nécessite une réévaluation des index : de nouveaux concepts ou relations peuvent apparaître dans l'ontologie ou bien certains concepts ou relations peuvent disparaître de l'ontologie. L'étude des impacts des évolutions de l'ontologie dans la mise à jour des index n'a pas été traitée dans cette thèse et font partie des perspectives à notre travail.

Les documents, une fois indexés sémantiquement, peuvent faire l'objet de recherches sémantiques sur les contenus. Cette recherche sémantique s'appuie sur la mesure de similarité sémantique entre les concepts de documents et ceux de la requête. Nous détaillons dans la section suivante nos solutions pour la recherche sémantique des documents.

3.4. Recherche de documents

La recherche de documents est le processus par lequel, moyennant un système informatique, l'utilisateur essaie de trouver et de sélectionner dans le corpus de documents ceux qui pourraient correspondre à ses besoins en information.

Dans cette section, nous rappelons le principe général des SRI (3.4.1). Nous abordons ensuite les notions fondamentales relatives à la prise en compte de la requête dans ces systèmes (3.4.2). Nous développons ensuite les fonctions de recherche de documents et les mesures de similarité conceptuelle (3.4.3). Nous terminons cette section par les détails sur les stratégies de reformulation de la requête de l'utilisateur (3.4.4).

3.4.1. Principe Général

La RI commence par l'expression du besoin de l'utilisateur sous forme d'une requête exprimée en langage libre ou bien par la sélection d'un document existant en vue de rechercher dans le corpus les documents sémantiquement similaire à ce dernier. Puis, après avoir comparé la requête par rapport aux documents, le système affiche les documents potentiellement pertinents à l'utilisateur ; celui-ci sélectionne les documents qu'il souhaite voir afficher. La Figure 19 présente me diagramme des cas d'utilisation correspondant.

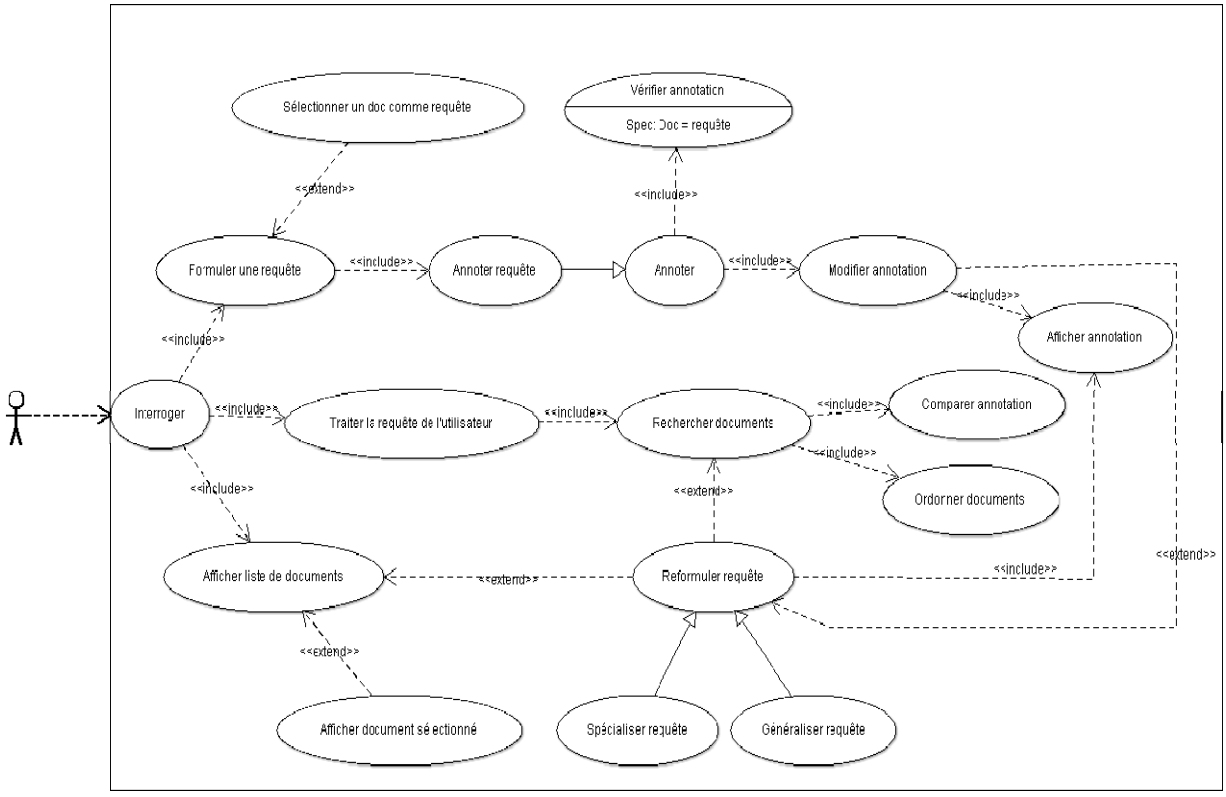

Figure 19 : Diagramme des cas d'utilisation de la RI (Laublet et al., 2009).

Ce diagramme montre le processus d'interrogation du SRI Dynamo. La requête saisie par l'utilisateur sera annotée automatiquement par le système. Cette annotation peut être modifiée par l'utilisateur si elle ne convient pas à ses besoins. Puis, le système affiche la liste des

documents potentiellement pertinents après avoir recherché ces derniers par comparaison de l'annotation de la requête avec celles des documents du corpus. Les détails de ce processus sont exposés dans les sous-sections suivantes.

3.4.2. Prise en compte de la requête

La requête, qu'elle soit du texte libre exprimant les besoins en information de l'utilisateur ou un document existant sélectionné, est annotée à l'aide des graphes d'instances de concepts de l'ontologie avant d'être comparée (évaluée) aux documents du corpus.

3.4.2.1. Formulation de la requête

La requête utilisateur peut être formulée de trois façons : soit, l'utilisateur saisit sa requête sous forme de textes libres, soit il navigue dans l'ontologie pour sélectionner des concepts (Hernandez, 2005) (Hubert et Mothe, 2009), soit il sélectionne un document existant qui sera traité comme une requête. Dans ce cas, le document sélectionné peut être un document déjà annoté ou non.

3.4.2.2. Annotation de requête

La requête saisie par l'utilisateur est annotée automatiquement en se basant sur les algorithmes de traitement automatique des langues naturelles. Le système essaye de trouver dans l'ontologie les éventuels concepts ou graphes de concepts qui peuvent être associés aux termes de la requête (Les mécanismes d'annotation par des graphes de concepts sont détaillés dans les sous-sections 3.2.3, 3.2.4 et 3.2.5). Le système propose alors à l'utilisateur ces graphes de concepts qui pourraient être associés à la requête. Puis, l'utilisateur peut modifier les propositions d'annotations si ces dernières ne correspondent pas à son besoin.

3.4.2.3. Évaluation de la requête

L'évaluation de la requête consiste en la mesure des similarités entre les annotations de la requête et celles des documents. Une requête donnée, aussi bien que les documents, peut être annotée avec plusieurs annotations. Le principe est de comparer tous les concepts ou graphes d'annotations de la requête avec ceux des documents et en gardant la similarité maximale. Les algorithmes d'évaluation de la requête que nous avons développés sont détaillés dans la section 3.4.3 suivante.

3.4.3. Recherche et similarité

La recherche de documents potentiellement pertinents dans un corpus par rapport à une requête passe par la mesure de similarité entre les graphes de concepts ou concepts isolés qui se trouvent dans la requête et celles des documents. En effet, dans un document comme dans une requête, plusieurs annotations ou graphes de concepts peuvent être associés. Le calcul de la similarité de graphes de concepts se base sur la mesure de la similarité entre les concepts.

Pourtant, l'une des problématiques actuelles de la RI sémantique est de trouver la meilleure mesure de similarité qui permette de prendre en compte la sémantique véhiculée par les termes trouvés dans les documents. Dans la littérature, les mesures de similarité proposées par (Resnik, 1995) et par (Jiang et Conrath, 1997) sont basées sur les statistiques d'utilisation des types de concepts sans tenir compte de la sémantique des concepts. La mesure proposée par (Wu et Palmer, 1994) tient compte de l'organisation en ontologie des concepts en particulier au travers de sa structure hiérarchique. Cependant, (Wu et Palmer, 1994) fait peu de

différence entre la relation de subsumption et la relation de partage d'un parent commun. De plus, dans (Wu et Palmer, 1994), deux concepts se trouvant dans deux sous-arbres taxonomiques différents de la structure hiérarchique des concepts d'une ontologie ont une valeur de similarité assez élevée alors que dans ce cas là, les deux concepts sont censés avoir des sémantiques très différentes dans certains cadres d'application (cf. section 4.2).

Nous proposons ci-après les différentes formules de mesure de similarité que nous avons développées pour répondre à ces problématiques. Puis nous comparons ces formules avec les formules de la littérature. Ensuite nous détaillons l'algorithme de la fonction de similarité entre document et requête en utilisant notre formule de mesure de similarité conceptuelle.

3.4.3.1. Mesure de similarité conceptuelle

Dans notre approche, nous considérons que les termes d'un vocabulaire peuvent être organisés sous forme d'arbre hiérarchique où les termes plus génériques sont rattachés avec les termes plus spécifiques comme des termes fils dans l'arbre. Ainsi, la mesure de similarité que nous proposons est inspirée du principe de l'arbre généalogique familial. La similarité entre deux concepts est assimilée à la proximité de deux membres de la famille (Nous appelons la fonction ProxiGénéa pour ***Proxi***mité ***Généa***logique).

Dans cette fonction, plus deux membres quelconques de la famille ont des ancêtres en commun, plus ils sont proches l'un de l'autre. Nous partons également de l'hypothèse que l'éloignement d'un membre de famille à partir d'un ancêtre commun influence sa distance par rapport aux autres membres de famille. La similarité entre deux concepts est donc une question de rapport entre le nombre d'ancêtres communs et la généalogie de ces concepts (l'ensemble des concepts de la racine jusqu'au concept).

Pour illustrer notre hypothèse, nous considérons l'extrait d'une ontologie présenté à la Figure *20* :

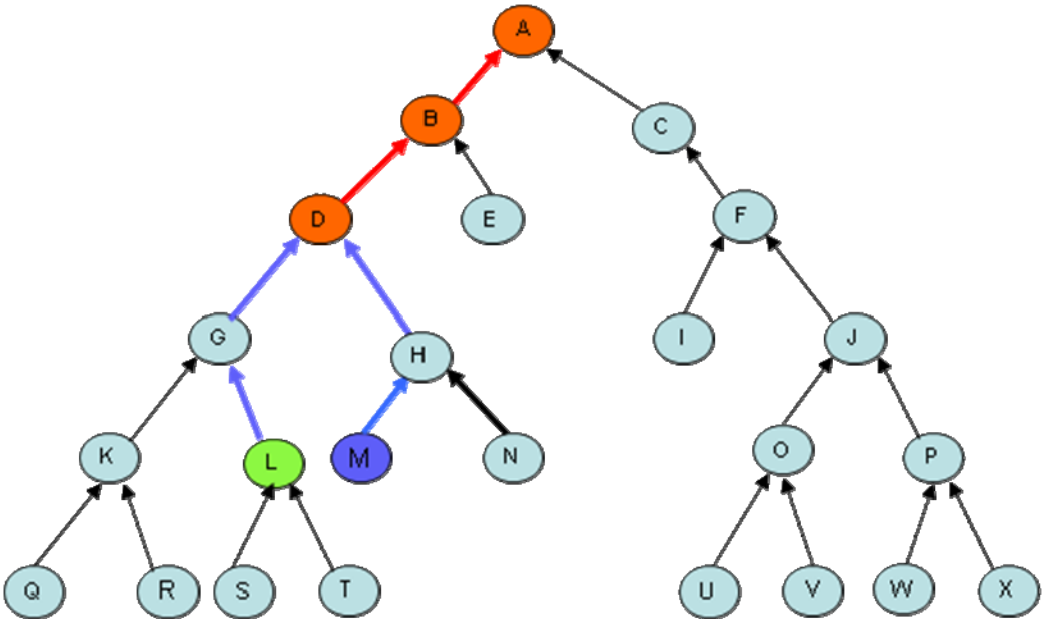

Figure 20 : Extrait d'ontologie.

- ***Gen(M)*** est l'ensemble des concepts qui entrent dans la généalogie du concept M, depuis la racine jusqu'à M.

$\text{Gen(M)} = \{\text{A, B, D, H, M}\}$

- ***Ancêtres (L, M)*** est l'ensemble des ancêtres communs des concepts L et M.

$$Ancêtres(L,M) = Gen(L) \cap Gen(M)$$
$$= \{A,B,D,G,L\} \cap \{A,B,D,H,M\}$$
$$= \{A,B,D\}$$

La fonction ***Card( )***, qui donne le cardinal d'un ensemble, exprime la position en profondeur d'un membre de famille. Par exemple, Card(Ancêtres(L,M))=3.

Pour bien appréhender la notion de similarité et de distance entre deux membres d'une famille dans un arbre généalogique donné, notre démarche consiste à considérer quatre points de vues qui aboutissent à diverses mesures de similarité. Nous allons détailler ci-après chacune de ces mesures de similarité avant de les comparer avec celle de Wu et Palmer.

3.4.3.2. Proximité avec l'ancêtre commun :

En partant du principe que :

- deux concepts sont plus proches s'ils sont plus proches des ancêtres communs,
- la similarité entre deux concepts est un rapport entre le nombre de leurs ancêtres communs et leurs généalogies,
- plus un membre de la famille s'éloigne du dernier ancêtre commun, plus il est distant par rapport à un autre membre de famille qui est plus près de cet ancêtre commun.

La proximité du concept L par rapport aux ancêtres communs avec le concept M est exprimée par le rapport :

$$\frac{Card\,(Ancêtres\,(L,M))}{Card\,(Gen(L))} \qquad \textbf{(23)}$$

De même, la proximité du concept M par rapport aux ancêtres communs avec le concept L est exprimée par le rapport :

$$\frac{Card\,(Ancêtres\,(L,M))}{Card\,(Gen(M))} \qquad \textbf{(24)}$$

La proximité d'un concept par rapport aux ancêtres communs avec un autre concept influence la similarité de l'un avec l'autre.

Nous proposons une première formule de similarité sémantique entre concepts :

$$Sim_{ProxiGénéa}(L,M) = \frac{Card\,(Ancêtres\,(L,M))}{Card\,(Gen(L))} * \frac{Card\,(Ancêtres\,(L,M))}{Card\,(Gen(M))}$$

$$\rightarrow\ Sim_{ProxiGénéa1}(L,M) = \frac{Card(Ancêtres(L,M))^2}{Card(Gen(L)) * Card(Gen(M))} \qquad \textbf{(25)}$$

Dans la suite, nous présentons des exemples relatifs aux différents cas de couples de concepts issus de la Figure 20. Nous indiquons le résultat obtenu par la formule de calcul.

Exemples :

- Cas de deux concepts provenant d'un ancêtre généralisant leurs parents respectifs (L et M) :

$$Sim(L,M) = \frac{Card(Ancêtres(L,M))^2}{Card(Gen(L)) * Card(Gen(M))} = \frac{3^2}{5*5} = 0{,}36$$

- Cas de deux concepts identiques :

$$Sim(L,L) = \frac{Card(Ancêtres(L,L))^2}{Card(Gen(L)) * Card(Gen(L))} = \frac{Card(Gen(L))^2}{Card(Gen(L))^2} = 1 \quad \text{car}$$

$$Ancêtres\,(L,L) = Gen(L) \cap Gen(L) = Gen(L) = \{A, B, D, G, L\}$$

- Cas de deux concepts liés par la relation « est un » (H et L) :

$$Sim(G,L) = \frac{Card(Ancêtres(G,L))^2}{Card(Gen(G)) * Card(Gen(L))} = \frac{Card(Gen(G))^2}{Card(Gen(G)) * Card(Gen(L))}$$

$$= \frac{Card\,(Gen(G))}{Card\,(Gen(L))} = \frac{4}{5} = 0{,}8$$

- Cas de deux concepts frères (K et L) :

$$Sim(K,L) = \frac{Card(Ancêtres(K,L))^2}{Card(Gen(K)) * Card(Gen(L))} = \frac{Card(Gen(G))^2}{Card(Gen(K)) * Card(Gen(L))} = 0{,}64$$

car $Ancêtres(K,L) = Gen(G)$

- Cas de deux concepts issus de deux branches taxonomiques différentes (Q et U) :

$$Sim(Q,U) = \frac{Card(Ancêtres(Q,U))^2}{Card(Gen(Q)) * Card(Gen(U))} = \frac{Card(Gen(A))^2}{Card(Gen(Q)) * Card(Gen(U))} = 0{,}02$$

car $Ancêtres(Q,U) = Gen(A)$

Les résultats que nous avons obtenu permettent d'illustrer que deux concepts liés par la relation de subsomption « est un » sont plus proches que ceux reliés par la relation de fratrie. De plus, deux concepts situés dans deux sous-arbres taxonomiques différents ont une similarité proche de zéro.

3.4.3.3. Rapport entre ancêtres communs

Soit $Card(Gen(L)) + Card(Gen(M))$ le nombre total de concepts qui entre dans la généalogie des deux concepts.

Sachant que $\text{Ancêtres(L, M)} \subseteq Gen(L)$ et $\text{Ancêtres(L, M)} \subseteq Gen(M)$, l'ensemble Ancêtres(L, M) figure deux fois dans l'ensemble des généalogies.

Nous considérons dans cette approche que la similarité entre deux concepts est un rapport entre ceux qui sont communs dans la généalogie de ces deux concepts. En d'autres termes, ceci revient à calculer, parmi ceux qui entrent dans la généalogie des deux concepts, combien font partie des ancêtres communs.

Dans ce cas, une deuxième mesure de similarité est définie par :

$$Sim_{\text{Pr}oxiGénéa\ 2}(L, M) = \frac{2 * Card\ (Ancêtres\ (L, M))}{Card\ (Gen(L)) + Card\ (Gen(M))} \qquad \textbf{(26)}$$

Cette formule a une analogie avec à la similarité conceptuelle de (Wu et Palmer, 1994) qui est :

$$Sim_{WP}(C1, C2) = \frac{2 * N3}{N1 + N2 + 2 * N3} \qquad \textbf{(27)}$$

Où N1 et N2 sont les nombres d'arcs qui séparent C1 et C2 de leur ascendant commun le plus spécifique C3, et N3 est le nombre d'arcs qui séparent C3 de la racine.

Etant donné que l'ensemble Ancêtres(L, M) figure deux fois dans l'ensemble des généalogies des deux concepts ($\text{Ancêtres(L, M)} \subseteq Gen(L)$ et que $\text{Ancêtres(L, M)} \subseteq Gen(M)$), nous pensons apporter une amélioration à cette formule en éliminant cette redondance d'ancêtres communs. Ainsi, notre deuxième mesure de similarité devient :

$$Sim_{\text{Pr}oxiGénéa\ 2}(L, M) = \frac{Card\ (Ancêtres\ (L, M))}{Card\ (Gen(L)) + Card\ (Gen(M)) - Card\ (Ancêtres\ (L, M))} \qquad \textbf{(28)}$$

qui peut aussi être exprimée par :

$$\rightarrow\ Sim_{\text{Pr}oxiGénéa\ 2}(L, M) = \frac{Card\ (Gen(L) \cap Gen(M))}{Card\ (Gen(L) \cup Gen(M))} \qquad \textbf{(29)}$$

La formule (28) correspond au rapport entre l'ensemble des ancêtres communs et l'ensemble des concepts de généalogie qui mènent vers les deux concepts à comparer.

L'ensemble des concepts de généalogie qui mènent vers les deux concepts L et M est $Gen(L) \cup Gen(M)$.

Avec
$Card(Gen(L) \cup Gen(M)) = Card(Gen(L)) + Card(Gen(M)) - Card(Gen(L) \cap Gen(M))$

où $Gen(L) \cap Gen(M) = Ancêtres(L, M)$.

Exprimée à la façon de Wu et Palmer, notre formule devient :

$$Sim_{ProxiGénéa2}(C1,C2) = \frac{N3}{N1+N2+N3} \qquad \textbf{(30)}$$

En partant de la formule initiale (28) qui est :

$$Sim_{ProxiGénéa\ 2}(L,M) = \frac{Card\ (Ancêtres\ (L,M))}{Card\ (Gen(L)) + Card\ (Gen(M)) - Card\ (Ancêtres\ (L,M))} \qquad \textbf{(31)}$$

Soit $\mathrm{Diff}_{\mathrm{Sym}}(\mathrm{L},\mathrm{M}) = \mathrm{Gen}(\mathrm{M})\Delta\mathrm{Gen}(\mathrm{L})$ **(32)** la différence symétrique de Gen(M) et de Gen(L), c'est à dire l'ensemble des concepts des généalogies des deux concepts qui ne font pas partie des ancêtres communs de L et de M. Avec $\mathrm{Gen(L)}\Delta\mathrm{Gen(M)} = \mathrm{Gen(L \setminus M)} \cup \mathrm{Gen(M \setminus L)}$

$Card(Diff_{Sym}(L,M)) = Card(Gen(L)) + Card(Gen(M)) - 2*Card(Ancêtres(L,M))$ est la distance entre les deux concepts L et M.

Dans notre cas

$Card\ (Diff\ (L,M)) = Card\ (Gen(L)) + Card\ (Gen(M)) - 2*Card\ (Ancêtres(L,M)) = 5+5-2*3 = 4$

Or, le dénominateur

$Card\ (Gen(L)) + Card\ (Gen(M)) - Card\ (Ancêtres(L,M)) =$

$Card\ (Gen(L)) + Card\ (Gen(M)) - Card\ (Ancêtres(L,M)) - Card\ (Ancêtres(L,M)) + Card\ (Ancêtres(L,M))$

$= Card(Gen(L)) + Card(Gen(M)) - 2*Card(Ancêtres(L,M)) + Card(Ancêtres(L,M))$

$= Card(Ancêtres(L,M)) + Card(Diff_{sym}(L,M))$

Ainsi la formule peut s'écrire:

➔ $$Sim_{ProxiGénéa2}(L,M) = \frac{Card(Ancêtres(L,M))}{Card(Ancêtres(L,M)) + Card(Diff_{sym}(L,M))} \qquad \textbf{(33)}$$

➔ $$Sim_{ProxiGénéa2}(L,M) = \frac{Card(Ancêtres(L,M))}{Card(Ancêtres(L,M)) \times \left(1 + \frac{Card(Diff_{sym}(L,M))}{Card(Ancêtres(L,M))}\right)}$$

Puis, en simplifiant par $Card\ (Ancêtres(L,M))$, on obtient :

$$\rightarrow \quad Sim_{ProxiGénéa2}(L,M) = \frac{1}{1+\left[\frac{Card(Diff_{sym}(L,M))}{Card(Ancêtres(L,M)}\right]} \quad \textbf{(34)}$$

En posant comme distance ProxiGénéa l'expression :

$$Dist_{ProxiGénéa}(L,M) = \frac{Card(Diff_{sym}(L,M))}{Card(Ancêtres(L,M))} \quad \textbf{(35)}$$

(qui est le rapport entre le cardinal de l'ensemble des concepts des généalogies des deux concepts qui ne font pas partie des ancêtres communs avec le cardinal des ancêtres communs des deux concepts), l'expression finale de notre deuxième mesure de similarité est donc :

$$Sim_{ProxiGénéa2}(L,M) = \frac{1}{1+Dist_{ProxiGénéaG}(L,M)} \quad \textbf{(36)}$$

Cette représentation de la formule facilitera sa comparaison avec une autre (40) formule que nous développons plus bas.

3.4.3.4. Inversion de distance par rapport aux ancêtres communs

Nous constatons que la similarité entre deux concepts est inversement proportionnelle à leurs distances par rapport aux ancêtres communs, c'est-à-dire, inversement proportionnelle au nombre de concepts qui ne font pas partie des ancêtres communs.

L'ensemble des concepts qui ne font pas partie des ancêtres communs de Gen(L) et de Gen(M) est donné par $\text{Gen(L} \setminus \text{M)} \cup \text{Gen(M} \setminus \text{L)} = \text{Gen(L)}\Delta\text{Gen(M)}$ qui est la différence symétrique de Gen(L) avec Gen(M).

Or, on sait que, $\text{Ancêtres(L , M)} \subseteq Gen(L)$ et $\text{Ancêtres(L, M)} \subseteq Gen(M)$

Donc, le nombre de concepts qui ne font pas partie des ancêtres communs est :

$$\text{Card(Gen(M)}\Delta\text{Gen(M))} = Card(Gen(L)) + Card(Gen(M)) - 2 * Card(Ancêtres(L,M))$$

Dans la littérature, (Lin, 1998) définit une classe de mesure de similarité basée sur la distance métrique entre deux concepts, dans laquelle, si la distance métrique entre deux objets est $dist(A,B)$, leur similarité est définie par :

$$Sim_{dist}(A,B) = \frac{1}{1+dist(A,B)} \quad \textbf{(37)}$$

En partant de cette idée, et en utilisant la différence symétrique que nous avons définie (32), nous obtenons:

$$Sim_{\text{Pr oxiGénéa 3}}(L,M) = \frac{1}{1+\left[Card\,(Gen(L)) + Card\,(Gen(M)) - 2*Card\,(Ancêtres\,(L,M))\right]} \quad \textbf{(38)}$$

3.4.3.5. Distance entre concepts

Afin d'illustrer cette approche par distance entre concepts, considérons les deux extraits d'ontologies suivants :

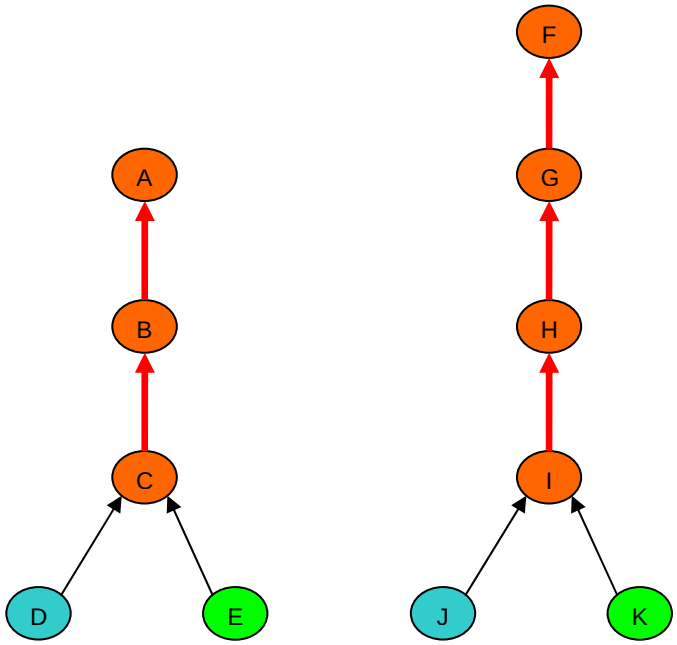

Figure 21 : Extraits d'ontologies montrant la proximité sémantique entre deux concepts.

En comparant les couples de concepts (D, E) et (J, K), nous voyons que le couple (J, K) a plus d'ancêtres communs que (D, E). Nous considérons dans cette approche que J et K sont plus similaires que D et E. En d'autres termes, D et E sont plus distants que J et K. Ainsi, moins deux concepts ont d'ancêtres communs, plus ils sont distants. Les autres formules (25) (28) prennent aussi en compte ce principe de similarité mais ici nous abordons l'idée de similarité à partir de la notion de distance entre deux concepts.

Contrairement à (Rada et al., 1989) qui expriment la distance entre deux concepts c_1 et c_2 comme le nombre minimum d'arcs à parcourir pour aller du concept c_1 au concept c_2, nous montrons que la notion de distance sémantique entre deux concepts ne dépend pas seulement du nombre de concepts qui séparent les deux concepts de l'ancêtre commun le plus spécifique, mais également du nombre de concepts ancêtres en commun.

Ainsi, nous considérons que la distance entre deux concepts se définit comme le rapport entre le nombre de concepts qui n'appartiennent pas aux ancêtres communs avec le nombre de concepts des ancêtres communs.

De ce fait, comme le nombre de concepts qui n'appartiennent pas aux ancêtres est :

$$Card(Gen(L)\Delta(Gen(M)) = Card(Gen(L)) + Card(Gen(M)) - 2 * Card(Ancêtres(L, M))$$

Nous avons $\frac{Card(Gen(D)\Delta Gen(E))}{Card(Ancêtres(D, E))} = \frac{2}{3} = 0{,}66$

Et $\frac{Card(Gen(J))\Delta Card(Gen(K))}{Card(Ancêtres(J, K))} = \frac{2}{4} = 0{,}50$

Nous voyons que les concepts D et E sont plus distants que les concepts J et K. En d'autres termes, les concepts J et K sont plus similaires que D et E. Ce calcul correspond bien à notre hypothèse.

En se basant sur la similarité de (Lin, 1998) suivante, $Sim_{dist}(A, B) = \frac{1}{1 + dist(A, B)}$ **(39)**

où dist(A, B) est la distance métrique entre le concept A et B.

Avec l'expression de distance que nous avons définie (35)

Nous arrivons à une nouvelle mesure de similarité :

$$Sim_{ProxiGénéa}(A, B) = \frac{1}{1 + Dist_{ProxiGénéa}(A, B)} \quad \textbf{(40)}$$

Nous pouvons constater que (40) n'est autre que (36).

Ainsi, en partant de deux points de vue différents, nous aboutissons sur une même formule de mesure de similarité entre deux concepts. Ce qui justifie le fondement de notre hypothèse sur la similarité conceptuelle.

3.4.3.6. Comparaison des mesures de similarités conceptuelles :

Ci-après le tableau 1 compare les différentes mesures de similarités ProxiGénéa par rapport à celle de Wu et Palmer. L'objectif est de montrer comment les différentes mesures de similarité prennent en compte les différents types de relation entre concepts et leurs positions relatives dans la hiérarchie de concepts. Chaque ligne d'un sous-tableau indique les valeurs de similarité entre les deux concepts indiqués. Le tableau du gauche affiche les similarités selon le type de relation entre concepts en comparant les mesures Wu et Palmer et ProxiGénéa1. Les lignes numérotées de 1 à 6 comprennent des couples de concepts frères. Les lignes 7 et 8 contiennent des concepts liés avec la relation de subsomption. Les lignes 9 à 12 concernent des concepts liés par la relation grand-père et petit-fils. Les concepts de la ligne 13 et 14 sont liés par la relation arrière-grand-père et arrière petit-fils. Les concepts des lignes 15 et 16 sont voisins. Les concepts des lignes 17 et 18 sont liés par la relation oncle-neveu. Les concepts de la ligne 19 sont liés par la relation arrière-arrière-grand-père. Les concepts de la ligne 21 à23 sont des concepts qui se trouvent dans deux sous-arbres taxonomiques différents. Le sous-tableau du milieu affiche les similarités triées suivant la valeur décroissante de la fonction Wu et Palmer et celui de droite affiche les similarités suivant la valeur décroissante de ProxiGénéa1.

N°	C1	C2	W&P	PG1
1	B	C	0,50	0,25
2	D	E	0,67	0,44
3	G	H	0,75	0,56
4	I	J	0,75	0,56
5	K	L	0,80	0,64
6	Q	R	0,83	0,69
7	K	G	0,89	0,80
8	K	Q	0,91	0,83
9	D	A	0,50	0,33
10	G	B	0,67	0,50
11	K	D	0,75	0,60
12	Q	G	0,80	0,67
13	Q	D	0,67	0,50
14	K	B	0,57	0,40
15	K	N	0,60	0,36
16	Q	T	0,67	0,44
17	G	E	0,67	0,50
18	K	H	0,67	0,45
19	Q	B	0,50	0,33
20	K	K	1,00	1,00
21	G	J	0,25	0,06
22	K	P	0,20	0,04
23	Q	X	0,17	0,03

N°	C1	C2	W&P	PG1
20	K	K	1,00	1,00
8	K	Q	0,91	0,83
7	K	G	0,89	0,80
6	Q	R	0,83	0,69
12	Q	G	0,80	0,67
5	K	L	0,80	0,64
11	K	D	0,75	0,60
4	I	J	0,75	0,56
3	G	H	0,75	0,56
16	Q	T	0,67	0,44
13	Q	D	0,67	0,50
18	K	H	0,67	0,45
17	G	E	0,67	0,50
10	G	B	0,67	0,50
2	D	E	0,67	0,44
15	K	N	0,60	0,36
14	K	B	0,57	0,40
19	Q	B	0,50	0,33
9	D	A	0,50	0,33
1	B	C	0,50	0,25
21	G	J	0,25	0,06
22	K	P	0,20	0,04
23	Q	X	0,17	0,03

N°	C1	C2	W&P	PG1
20	K	K	1,00	1,00
8	K	Q	0,91	0,83
7	K	G	0,89	0,80
6	Q	R	0,83	0,69
12	Q	G	0,80	0,67
5	K	L	0,80	0,64
11	K	D	0,75	0,60
3	G	H	0,75	0,56
4	I	J	0,75	0,56
10	G	B	0,67	0,50
17	G	E	0,67	0,50
13	Q	D	0,67	0,50
18	K	H	0,67	0,45
2	D	E	0,67	0,44
16	Q	T	0,67	0,44
14	K	B	0,57	0,40
15	K	N	0,60	0,36
9	D	A	0,50	0,33
19	Q	B	0,50	0,33
1	B	C	0,50	0,25
21	G	J	0,25	0,06
22	K	P	0,20	0,04
23	Q	X	0,17	0,03

Tableau 1 : Comparaison des mesures de similarités sémantiques.

L'analyse de ces tableaux montre que ProxyGénéa1 et Wu et Palmer renvoient des résultats des mesures de similarité dans des ordres différents car ProxiGénéa1 tient en compte la structure hiérarchique de l'arbre des concepts. Nous constatons que plusieurs points communs existent entre les mesures ProxiGénéa1 et Wu et palmer. Toutes les deux privilégient la relation de subsomption par rapport à celle de la fratrie. De même, elles privilégient la relation de fratrie par rapport à la relation grand-père-petit-fils. Il est aussi constaté que pour deux concepts liés par les mêmes types de relations et qui sont dans le même sous-arbre taxonomique, ceux qui se trouvent en bas de la généalogie sont plus similaires. Par contre, deux concepts qui se trouvent dans deux sous-arbres taxonomiques différents sont plus distants lorsqu'ils se trouvent en bas de la hiérarchie.

Le tableau 2 présente les résultats obtenus par les autres fonctions que nous proposons. Les résultats des mesures sont triés par rapport aux résultats de la mesure ProxiGénéa3 dans le sous tableau de droite, et par celle Wu et Palmer dans celui du centre. Le sous tableau de gauche est trié en fonction des types de relation entre concepts.

N°	C1	C2	W&P	
1	B	C	0,50	0,33
2	D	E	0,67	0,33
3	G	H	0,75	0,33
4	I	J	0,75	0,33
5	K	L	0,80	0,33
6	Q	R	0,83	0,33
7	K	G	0,89	0,50
8	K	Q	0,91	0,50
9	D	A	0,50	0,33
10	G	B	0,67	0,33
11	K	D	0,75	0,33
12	Q	G	0,80	0,33
13	Q	D	0,67	0,25
14	K	B	0,57	0,25
15	K	N	0,60	0,20
16	Q	T	0,67	0,20
17	G	E	0,67	0,33
18	K	H	0,67	0,25
19	Q	B	0,50	0,20
20	K	K	1,00	1,00
21	G	J	0,25	0,14
22	K	P	0,20	0,11
23	Q	X	0,17	0,09

N°	C1	C2	W&P	PG3
20	K	K	1,00	1,00
8	K	Q	0,91	0,50
7	K	G	0,89	0,50
6	Q	R	0,83	0,33
5	K	L	0,80	0,33
12	Q	G	0,80	0,33
3	G	H	0,75	0,33
4	I	J	0,75	0,33
11	K	D	0,75	0,33
2	D	E	0,67	0,33
10	G	B	0,67	0,33
17	G	E	0,67	0,33
18	K	H	0,67	0,25
13	Q	D	0,67	0,25
16	Q	T	0,67	0,20
15	K	N	0,60	0,20
14	K	B	0,57	0,25
1	B	C	0,50	0,33
9	D	A	0,50	0,33
19	Q	B	0,50	0,20
21	G	J	0,25	0,14
22	K	P	0,20	0,11
23	Q	X	0,17	0,09

N°	C1	C2	W&P	PG3
20	K	K	1,00	1,00
8	K	Q	0,91	0,50
7	K	G	0,89	0,50
6	Q	R	0,83	0,33
5	K	L	0,80	0,33
12	Q	G	0,80	0,33
3	G	H	0,75	0,33
4	I	J	0,75	0,33
11	K	D	0,75	0,33
2	D	E	0,67	0,33
10	G	B	0,67	0,33
17	G	E	0,67	0,33
1	B	C	0,50	0,33
9	D	A	0,50	0,33
18	K	H	0,67	0,25
13	Q	D	0,67	0,25
14	K	B	0,57	0,25
16	Q	T	0,67	0,20
15	K	N	0,60	0,20
19	Q	B	0,50	0,20
21	G	J	0,25	0,14
22	K	P	0,20	0,11
23	Q	X	0,17	0,09

Tableau 2 : Comparaison des mesures de similarité sémantique.

L'analyse du tableau 2 montre que ProxyGénéa3 et Wu et Palmer renvoient des résultats des mesures de similarité dans des ordres différents.

Nous pouvons aussi observer à partir des deux tableaux 1 et 2 que toutes les mesures de similarités donnent plus de proximité à un lien parent-enfant (cas de K et G) qu'à un lien de partage de parent (cas de K et L). Cependant, nous pouvons constater que la mesure de Wu et Palmer distingue peu ces deux types de liens car le score du lien de parenté est de 0,88 alors que celui du lien de partage de parent est de 0,80. De plus, du moins dans notre cadre d'application (Projet Dynamo), cette mesure de similarité n'est pas très adaptée car elle ne met pas en évidence le fait que, nécessairement, deux concepts issus de deux branches taxonomiques différentes sont sémantiquement très distants. Pour le cas de (Q,X), qui sont deux concepts très distants et issus de deux branches taxonomiques différentes, Wu et Palmer donne la valeur de similarité 0,17. Nous pouvons constater que les mesures de similarité ***ProxiGénéa*** (1, 2 et 3) mettent toutes en évidence cette propriété. Par exemple, le même couple a pour valeur de similarité 0,03 pour ***ProxiGénéa1*** et 0,09 pour ***ProxiGénéa 2 et 3***.

La mesure ***ProxiGénéa3*** a la particularité de donner la même valeur de similarité 0,33 pour tout couple de concepts frères, la valeur de similarité 0,50 pour tout couple de concepts liés par une relation de subsomption, la valeur 0,20 pour tout couple de concepts voisins, la valeur 0,25 pour tout couple de concepts liés par la relation arrière-grand-père et la valeur de similarité 0,33 entre un grand-père et son petit-fils.

Nous pouvons constater que les valeurs de la formule ***ProxiGénéa2*** se placent toujours entre celles de Wu et Palmer et de ***ProxiGénéa1.*** De plus, les mesures ***ProxiGénéa2*** et Wu et Palmer renvoient toutes les deux les résultats dans le même ordre. Cependant, les valeurs de similarité sont différentes, ce qui peut avoir une incidence sur des systèmes qui ne retiendraient que les documents ayant des scores dépassant un seuil déterminé.

En ce qui concerne les mesures de similarité de Lin et de Resnik, nous ne sommes pas entièrement d'accord sur le fait que la notion de similarité de deux concepts d'une même ontologie doive être variable, en fonction des fréquences d'apparition de ces concepts dans les documents du corpus. Selon notre approche, une similarité sémantique entre deux concepts est fonction de l'essence même de ces concepts mais pas de la statistique de leurs apparitions dans les documents. En d'autres termes, ce ne sont ni le style de rédaction des auteurs ni le nombre de leurs rédactions qui affectent la similarité sémantique entre deux entrées d'un vocabulaire.

Pour aboutir à une RI sémantique, ces différentes mesures de similarité conceptuelle sont utilisées dans une mesure de similarité de graphes de concepts que nous présentons dans la section suivante.

3.4.3.7. Similarité de graphes de concepts

Une annotation donnée est considérée comme un ensemble soit de concepts en relation (ou un graphe de concepts) soit de concepts isolés. Un document ou une requête peut être annoté par un ou plusieurs graphes de concepts ou concepts isolés.

Pour calculer la similarité entre deux annotations (une annotation de requête et une annotation de document), nous mesurons les similarités entre les concepts de même type. Nous combinons ensuite les résultats obtenus en respectant l'ordre d'importance entre les types de concepts car tous les concepts d'une annotation n'ont pas la même importance.

Ainsi, pour prendre en compte la notion d'importance relative entre les différents types de concepts, nous calculons la similarité entre deux annotations comme la moyenne pondérée des similarités entre les concepts qui les composent. Les valeurs des coefficients sont à fixer pour chaque application.

$$Similarité(A_{req}, A_{doc}) = \frac{\sum_{i=0}^{taille(A_{req})} Coef[i] * ProxiGénéa(A_{req}[i], A_{doc}[i])}{\sum_{i=0}^{taille(A_{req})} Coef[i]} \quad \textbf{(41)}$$

Étant donné qu'un document et une requête peuvent avoir plusieurs annotations (une liste de graphes), la similarité entre deux listes de graphes peut être calculée de quatre façons :

- le maximum des similarités d'annotations, dans le cas où le fait d'avoir trouvé une meilleure similarité suffirait pour valider la similarité entre deux listes de graphes,
- la combinaison linéaire des similarités d'annotation, si l'on veut considérer toutes les annotations et combiner leurs similarités pour faire un classement de tous les documents résultats. Dans ce cas, soit le nombre de documents à trouver est limité à un nombre fixé à l'avance, soit le seuil de similarité sera calculé en fonction des ensembles des scores des documents du corpus,
- la moyenne pondérée des similarités d'annotation, pour prendre en compte toutes les annotations d'un document mais également pour donner de l'importance relative à chaque type d'annotation. Une annotation donnée peut être axée sur une branche taxonomique de l'ontologie de laquelle les concepts sont issus,
- le produit des similarités d'annotations, dans le cas où une mauvaise ou faible similarité entre deux annotations suffirait pour ne pas valider la similarité de deux listes de graphes.

Le choix de l'un de ces modes de calcul dépend de l'application cible dans laquelle le calcul de similarité est implémenté.

Nous proposons ci-après l'algorithme général du calcul de similarité entre une requête et un document.

```
Entrées : Requête, Corpus
Sorties : valeur réelle indiquant la similarité entre requête et document
Début
    Pour chaque document du corpus Faire
        Extraire les Annotations_du_document
        Pour chaque Annotation_du_document Faire
            Extraire les Concepts_document
            Classifier les Concepts par Type
            Extraire les Annotations_de_la_requête
            Pour chaque Annotation_de_la_requête Faire
                Extraire les Concepts_requête
                Classifier les Concepts par Type
                Calculer la similarité entre les Concepts_requête et les
                Concepts_document de même type
                Calculer Similarité (Annotations_de_la_requête,
                Annotation_du_document)
                Garder la valeur de similarité
            FinPour
            Garder la valeur de similarité
        FinPour
        Calculer Similarité (Requête, Document)
```

```
            /* Le calcul de similarité implémentera l'une des quatre méthodes de calcul de
            similarité entre liste de graphes (Max, Combinaison linéaire, Moyenne, Produit)*/
            Renvoyer le résultat
        FinPour
Fin
```

Algorithme 5 : Fonction de similarité entre document et requête.

La similarité de graphes de concepts permet de comparer sémantiquement les requêtes avec les documents. Cependant, le résultat issu de la comparaison de l'annotation (graphe de concepts) de requête avec celle des documents peut ne pas satisfaire l'utilisateur. En effet, le résultat obtenu peut renvoyer soit trop de documents, soit peu de documents, soit des documents qui ne correspondent pas à l'attente de l'utilisateur. Dans ce cas là, la reformulation de la requête s'avère nécessaire en vue d'obtenir d'autres résultats plus satisfaisants. Nous proposons ci-après nos stratégies de reformulation de requête.

3.4.4. Reformulation de requêtes

La reformulation de requêtes peut être automatique ou manuelle suivant que le système ou l'utilisateur l'effectue. Ainsi, la reformulation de requêtes peut intervenir en particulier à deux niveaux :

- lors de la recherche initiale, si aucun document n'est retrouvé, le système réalise la reformulation,
- si le lecteur n'est pas satisfait de la réponse du système après une première recherche, il peut reformuler sa requête.

Pour reformuler une requête, nous proposons quatre modes de reformulation : la généralisation, la spécialisation, la reformulation hybride qui combine sur une seule requête la généralisation et la spécialisation de la requête et enfin la prise en compte des concepts voisins.

3.4.4.1. Généralisation de la requête

Dans le cas où le nombre de documents trouvés est insuffisant, le système ou l'utilisateur peut procéder à la généralisation de la requête en prenant comme nouveaux concepts annotant la requête le concept parent (plus général) de chaque concept de la requête.

3.4.4.2. Spécialisation de la requête

Dans le cas contraire, si le nombre de documents restitués est très important, le système ou l'utilisateur peut procéder à la spécialisation de la requête en prenant comme nouveaux concepts annotant la requête un ou plusieurs concepts spécifiques subsumés par chaque concept de la requête.

3.4.4.3. Reformulation hybride (spécialisation & généralisation partielle)

La reformulation hybride est une façon de reformuler la requête en vue d'obtenir plus de précision sur les résultats. Cette reformulation se fait en spécialisant certains types de concepts et en en généralisant d'autres.

3.4.4.4. Prise en compte des concepts voisins

La prise en compte des concepts voisins est une façon d'étendre le résultat d'une recherche. En effet, les documents annotés par les concepts voisins d'un concept choisi par l'utilisateur peuvent l'intéresser. Sa différence avec la généralisation est qu'elle peut prendre en compte d'autres concepts.

3.4.5. Conclusion

La recherche de documents se fait par la comparaison des graphes de concepts ou concepts isolés annotant les documents avec ceux de la requête.

Nous avons développé une fonction générique de similarité de graphes de concepts entre documents et requête. Cette fonction est basée sur la mesure de similarité conceptuelle appelée ***ProxiGénéa*** inspirée de la proximité généalogique des membres d'une famille. Comme la structure des arbres généalogiques est analogue aux hiérarchies de concepts que l'on trouve dans les ontologies, ProxiGénéa est adaptée pour mesurer les similarités sémantiques entre les concepts. Ceci est du au fait que les différentes propriétés des hiérarchies de concepts ont été prises en compte. Les expérimentations que nous détaillons dans la section 4.2.7 montrent que ProxiGénéa améliore la qualité de la RI en termes de rappel et précision. De plus, par rapport à (Wu et Palmer, 1994), ProxiGénéa respecte mieux les propriétés de la proximité qui relient les concepts dans la hiérarchie.

En plus de la fonction de similarité, nous avons aussi proposé une stratégie de recherche de documents qui consiste d'une part en la généralisation ou la spécialisation des concepts selon que le résultat renvoie peu ou beaucoup de résultats, et d'autre part prenant en compte des concepts voisins pour développer l'étendue de la requête.

La fonction de similarité et la stratégie de RI que nous proposons ont été mises en œuvre dans le projet Dynamo que nous détaillons dans la section 4.2.

Le fondement de la théorie c'est la pratique.

[Mao Tsé-Toung]

Chapitre 4. APPLICATION / EVALUATION

Le modèle de représentation multi-facette des documents que nous avons proposé dans la section 3.1.1 a été instancié dans divers domaines d'application dont :

- l'apprentissage en ligne, dans lequel nous avons développé un prototype d'outil d'apprentissage en ligne (section 4.1),
- le dépannage automobile, dans lequel nous avons contribué à la réalisation d'un outil d'indexation dynamique et sémantique des fiches techniques en vue de leur recherche sémantique (section 4.2),
- le domaine de la musique Braille, dans lequel nous avons contribué à la conception et la réalisation d'un langage de représentation des documents musicaux Braille (section 4.3).

4.1. Prototype d'outils d'apprentissage en ligne

La finalité de toute RI est d'apprendre de nouvelles connaissances. Ainsi, l'apprentissage en ligne est un cas particulier de RI d'accès à l'information. Dans ce cas, les documents à rechercher par l'utilisateur sont des objets pédagogiques.

4.1.1. Objets pédagogiques

L'ensemble des objets pédagogiques mis au format package SCORM sont stockés dans un répertoire commun qui servira de lieu de dépôt et de recherche des objets pédagogiques pour leur réutilisation.

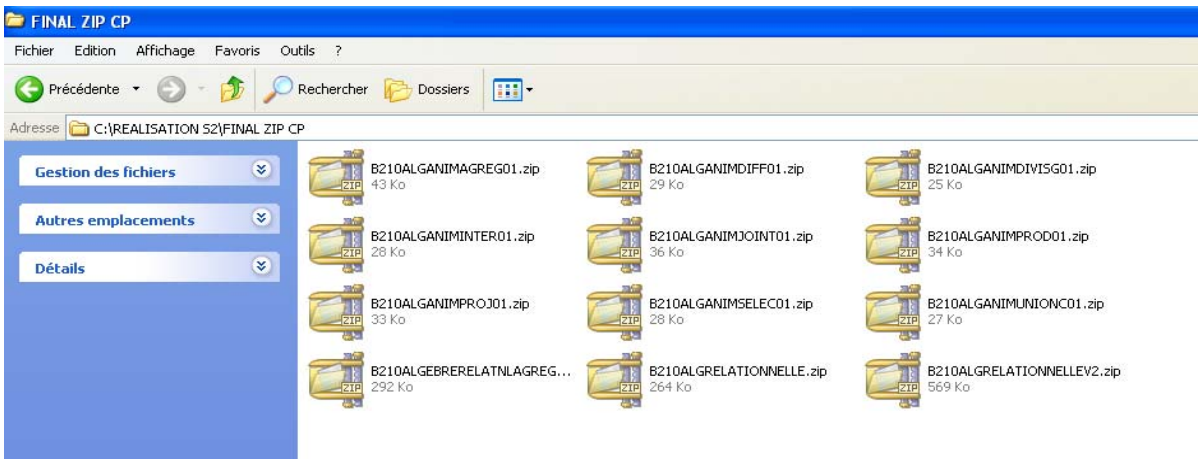

Figure 22 : Des packages d'objets pédagogiques prêts pour réutilisation.

Chacun de ces objets pédagogiques est structuré suivant le format décrit dans les Figure 23 et Figure 24 ci-dessous.

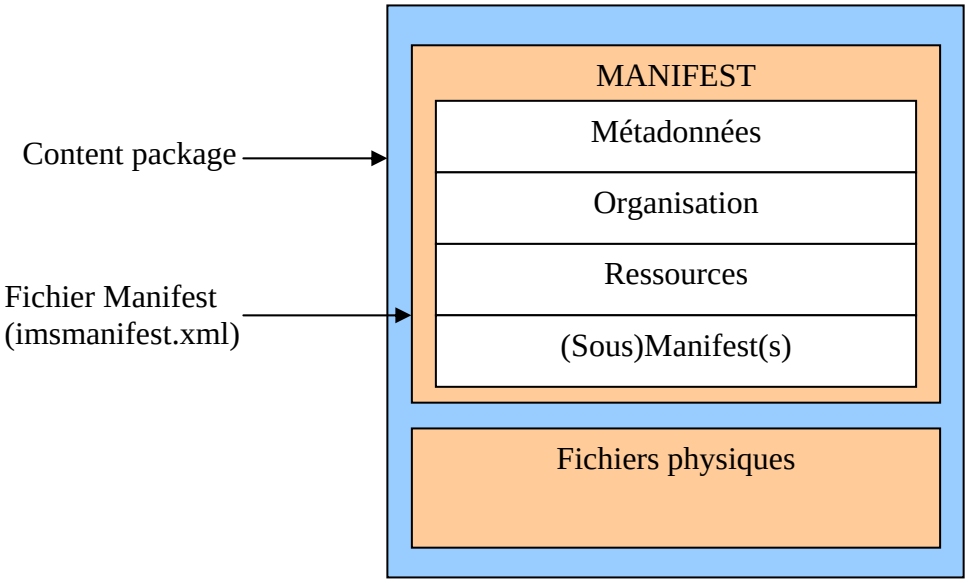

Figure 23 : Schéma conceptuel d'un package SCORM.

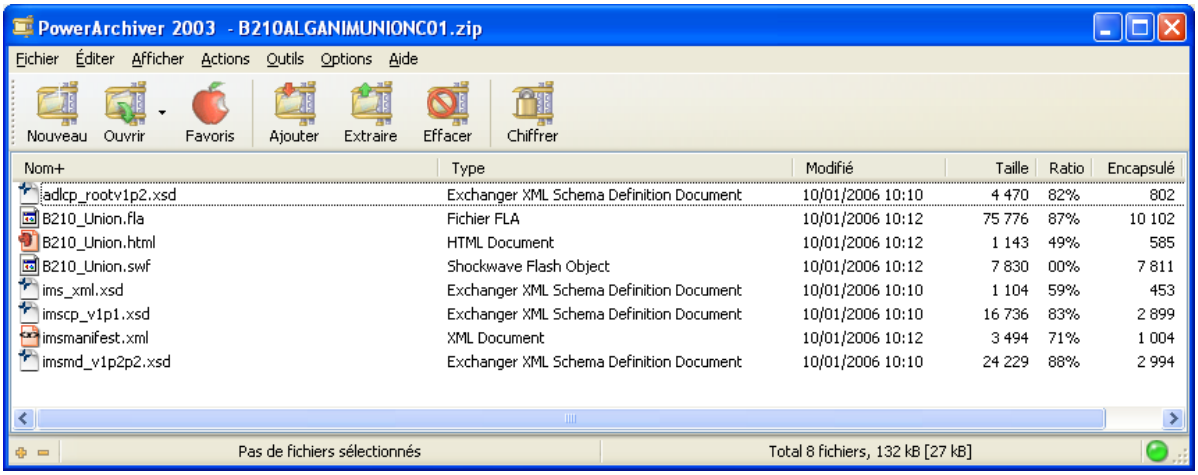

Figure 24 : Exemple de Package d'objets pédagogiques conforme à la norme SCORM.

La Figure 23 représente le contenu d'un objet pédagogique mis au format de package SCORM. L'ensemble des composants élémentaires qui forment un objet pédagogique ainsi que les fichiers de description sont regroupés dans un fichier compressé pour former un package d'objets pédagogiques.

Les métadonnées qui se trouvent dans le fichier imsmanifest.xml sont issues du profil d'application EMIAGE comme illustré dans la Figure 25.

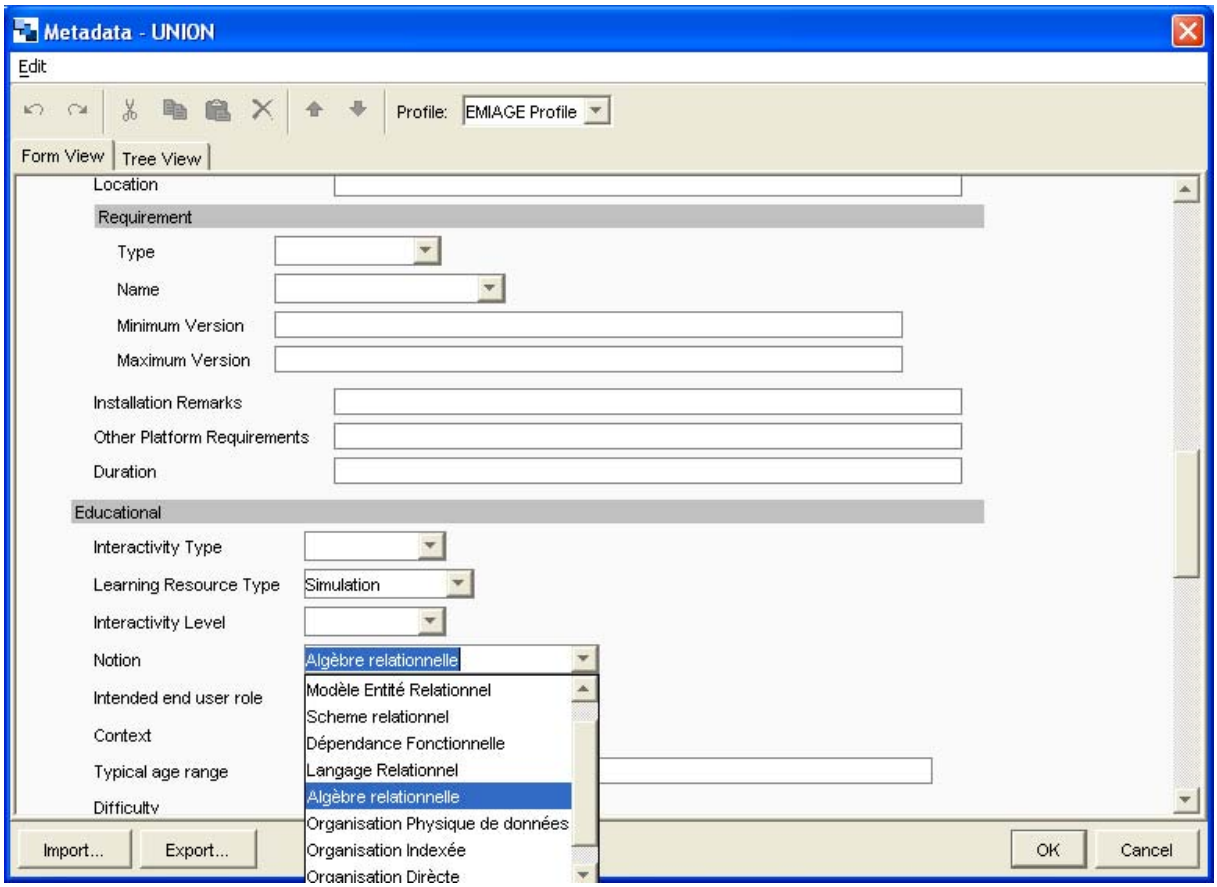

Figure 25 : Edition d'une métadonnée suivant le profil EMIAGE.

4.1.2. Indexation des objets pédagogiques

Nous présentons dans cette section, un exemple de représentation d'un objet pédagogique préparé pour le module Base de Données Relationnelles des L3 préparant la MIAGE (Maîtrise Informatique Appliqué à la Gestion des Entreprises) en formation à distance. L'objet pédagogique représenté est un exercice portant sur l'indexation de fichiers de données. Sa représentation à partir de notre modèle à facettes est illustrée afin de montrer l'intérêt d'une telle représentation lors de son intégration dans le système d'apprentissage (que ce soit pour une utilisation par l'enseignant ou par l'apprenant).

Par rapport à sa structure, cet objet est constitué de trois objets élémentaires : deux images de b-arbres au format jpg et un énoncé. La décomposition de l'objet à partir de sa structure permet d'envisager la réutilisabilité de chacun des éléments.

Les composants élémentaires ainsi que l'objet pédagogique qu'ils constituent sont indexés par rapport aux métadonnées LOM. Par exemple, la métadonnée « droit » des deux images est définie avec la valeur « public ». Ceci permet à n'importe quel enseignant ou apprenant d'y accéder et de l'utiliser. Par contre, pour l'énoncé cette métadonnée est définie comme propre aux personnes de la formation. Ceci implique que l'énoncé ainsi que l'objet pédagogique exercice ne pourra être réutilisé que par des enseignants de la formation et consulté par des apprenants inscrits. La métadonnée pédagogie-niveau relative à l'exercice est fixée à initiation. Ceci indique que l'exercice s'adresse à des étudiants n'ayant jamais étudié l'indexation de fichiers et qu'il pourra être réutilisé dans le cadre d'autres modules s'adressant à un public ayant le même niveau.

L'objet pédagogique est également représenté à partir des notions qu'il aborde. L'ontologie du thème des bases de données est présentée dans la Figure 26. Les concepts sont représentés par des rectangles contenant les différents labels ou termes permettant de définir les notions, les flèches légendées représentent les relations sémantiques entre concepts. Dans l'extrait proposé, les concepts surlignés représentent les notions à aborder dans le cadre du module « Base de Données Relationnelles ».

Une des images représente un b-arbre +, et l'autre un b-arbre *. Les deux images sont donc, dans notre modèle, représentées à partir de ces deux concepts de l'ontologie de des bases de données. L'exercice quant à lui aborde les notions d'organisation indexée des données à partir de b-arbres tout en insistant sur le temps d'accès à la base. Ces trois concepts de l'ontologie sont donc utilisés pour l'indexer. L'avantage de spécifier ces concepts est qu'un enseignant pourra réutiliser cet exercice lorsqu'il voudra travailler sur les notions précédemment cités.

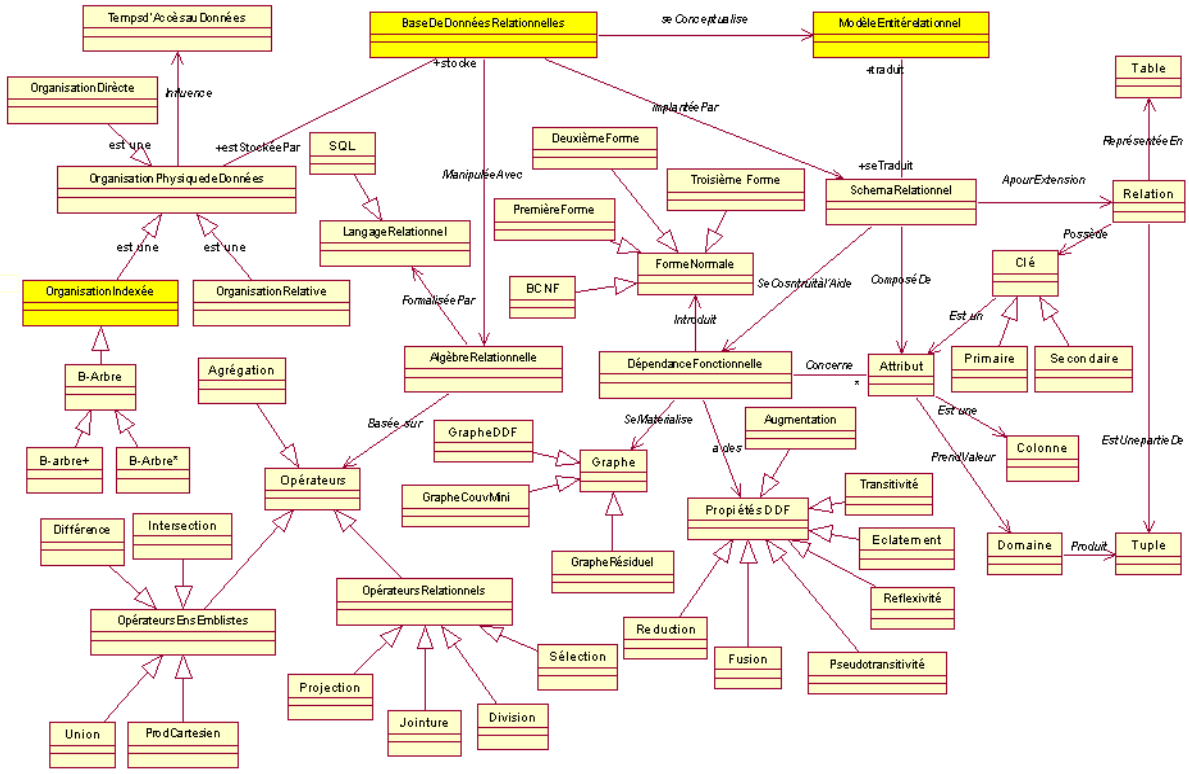

Figure 26 : Extrait de l'ontologie du thème des bases de données.

Dans le cadre de son utilisation dans un scénario pédagogique, l'exercice proposé pourra être utilisé à différents niveaux. Si l'on considère la pédagogie de Gagné (Lebrun, 2002), il pourra être utilisé pour provoquer la performance, fournir la rétroaction et évaluer la performance. Son utilisation est indiquée dans le diagramme d'activités présenté dans la Figure 28 au niveau des actes en gras (4 et 5).

4.1.3. Scénario d'apprentissage

Pour illustrer la notion de scénario d'apprentissage, nous considérons l'exemple présenté aux Figure 27 et Figure 28 (Hernandez et al, 2006). Dans cet exemple, l'enseignant présente les notions à appréhender aux étudiants. Ces derniers lisent les chapitres correspondants et demandent des explications à l'enseignant. L'enseignant donne des explications à l'aide des exemples et attend une nouvelle réaction des étudiants. Il donne éventuellement des suppléments d'explication au cas où les étudiants en demandent.

Puis, le délai écoulé, l'enseignant donne l'exercice que les étudiants doivent traiter. Ensuite, l'enseignant corrige des exercices tout en fournissant des explications précises sur chaque point non maîtrisé par l'étudiant, par courrier électronique, dialogue en direct ou sur un forum. Afin d'assurer la compréhension et la mémorisation des notions à appréhender, des exercices d'évaluation et d'auto-évaluation sont fournis aux étudiants. Finalement, l'enseignant donne un résumé des points à retenir sur la notion étudiée.

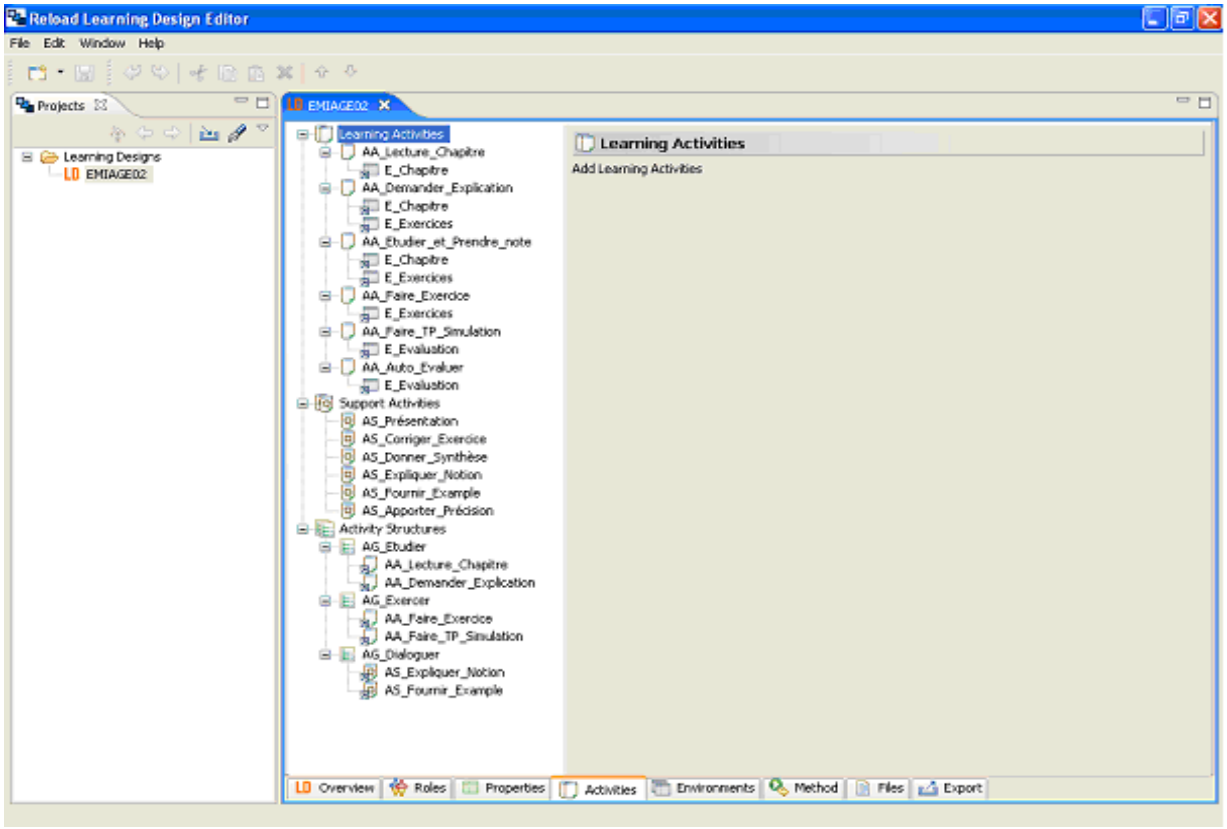

Figure 27 : Fenêtre d'édition des activités pédagogiques.

Cette figure montre la fenêtre d'édition des activités d'apprentissage avec le logiciel Reload LD-Editor[7].

[7] http://www.reload.ac.uk/ldeditor.html

Le séquencement de ces activités est détaillé dans le diagramme d'activité du scénario pédagogique présenté dans la Figure 28.

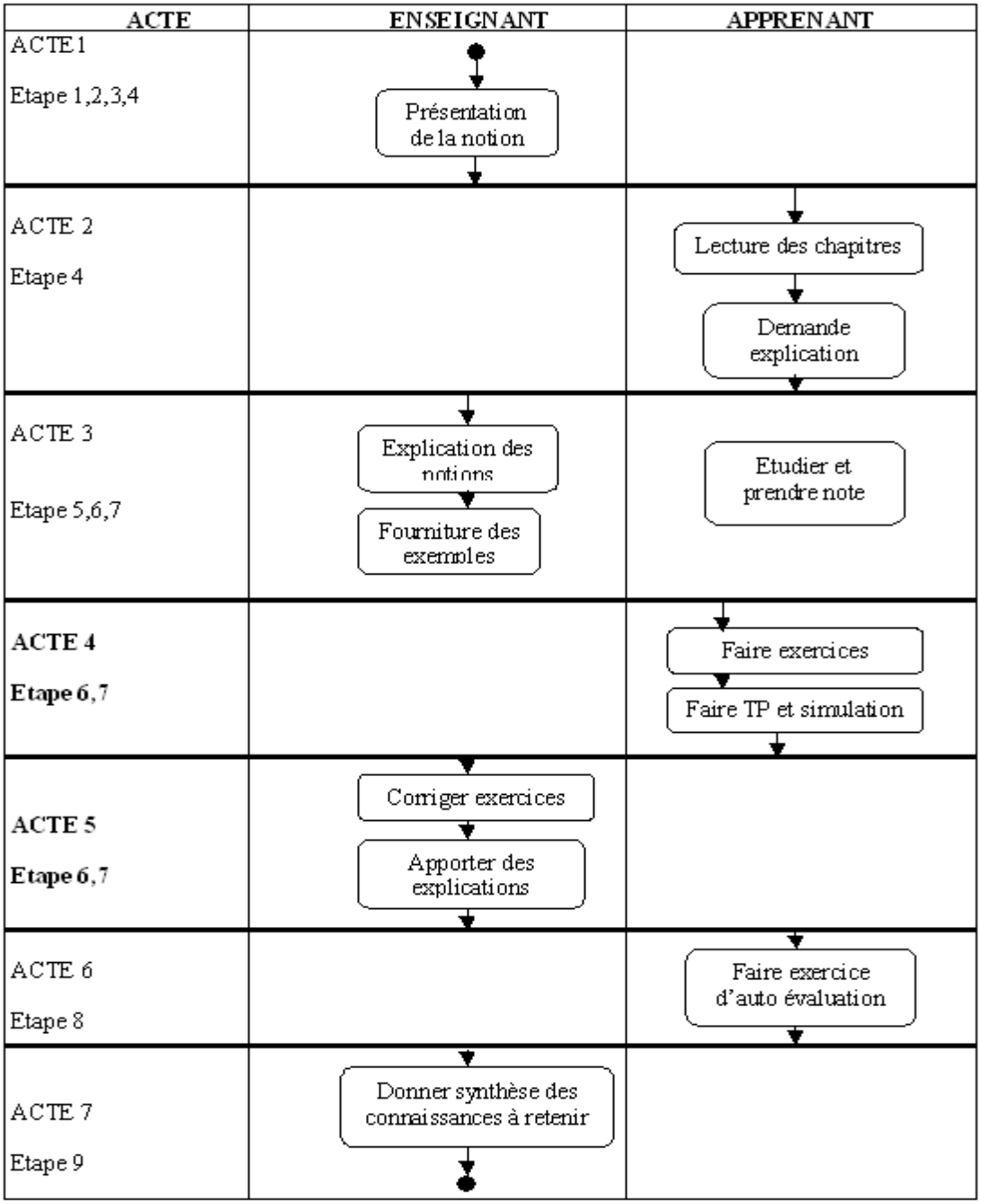

Figure 28 : Diagramme d'activité du scénario pédagogique intégrant l'exercice.

4.1.4. Prototype

Nous avons développé un prototype de système d'apprentissage baptisé PALM pour **P**lateforme d'**A**pprentissage en **L**igne **M**ultimédia. Cet outil permet d'importer des cours au format IMS-LD, précédemment créés à l'aide de Reload LD-Editor. Nous présentons ci-après quelques copies d'écran montrant le fonctionnement de PALM en s'appuyant sur des activités pédagogiques du scénario d'apprentissage décrit dans la Figure 28. Chacune de ces activités pédagogiques utilise et affiche les objets pédagogiques qui leurs sont assignés.

La Figure 29 ci-dessous représente l'exécution de l'*Activité d'Apprentissage Etudier et Prendre note* tandis que la Figure 30 montre un exemple d'utilisation d'objets pédagogiques de type animation pendant l'exécution de l'activité pédagogique *Auto-évaluation*.

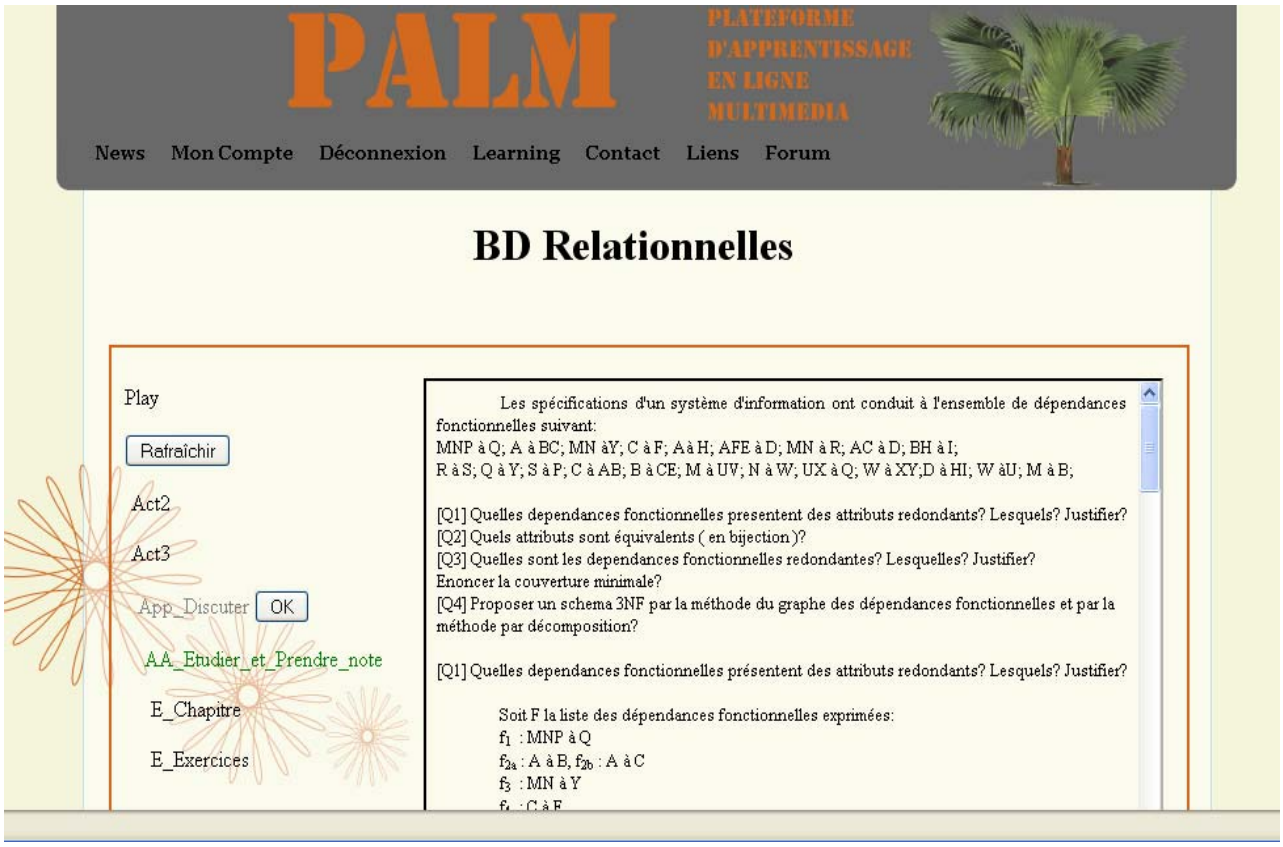

Figure 29 : Exécution de l'activité d'apprentissage étudier et prendre note.

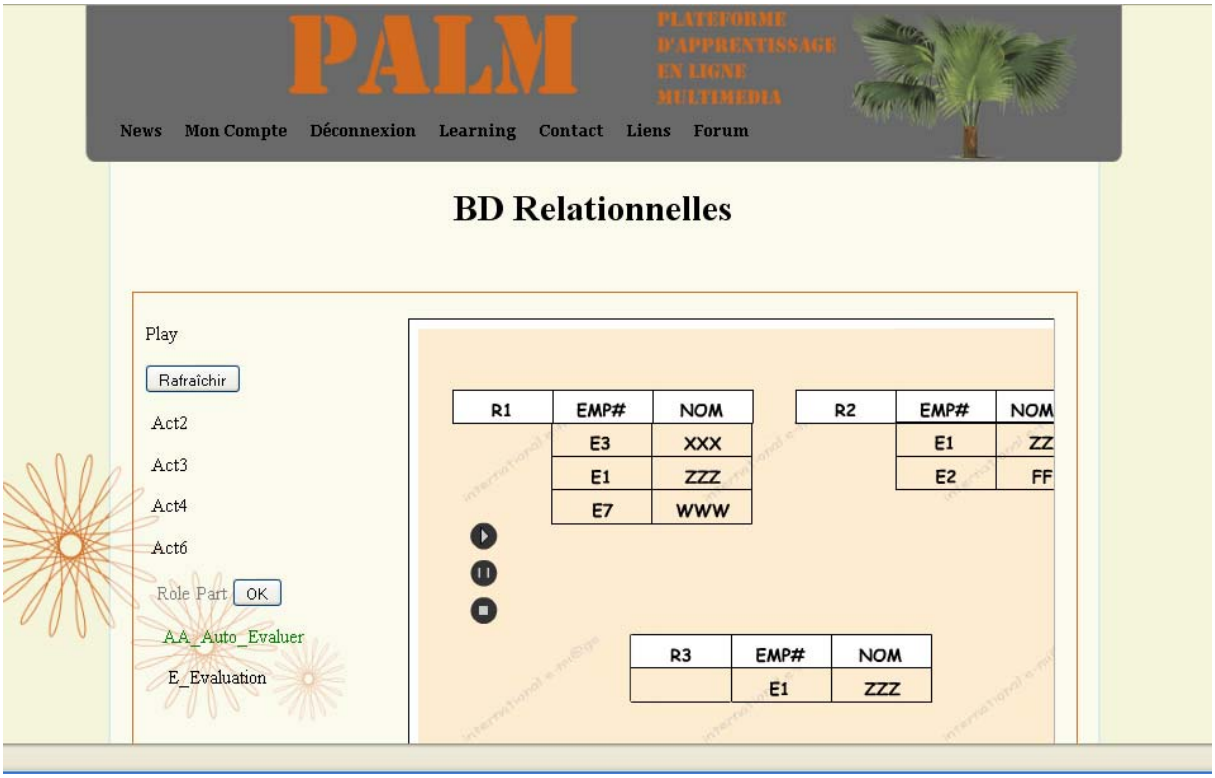

Figure 30 : Illustration d'une notion à l'aide d'un objet pédagogique de type animation.

Cette figure montre l'affichage d'une animation, illustrant la notion d'*agrégation*, qui fait partie de l'objet pédagogique *LO_Chapitre* utilisé dans l'activité *AA_Lecture_Chapitre* (cf. Figure 27).

4.1.5. Recherche d'objets pédagogiques

La recherche d'objets pédagogiques se fait à l'aide de l'interface dédiée. L'utilisateur saisit sa requête sous forme de termes correspondants aux concepts de type *Notion* de l'ontologie de domaine et le système affiche non seulement les différentes notions associées à la requête mais également les différents cours qui utilisent les objets pédagogiques indexés par la notion spécifiée.

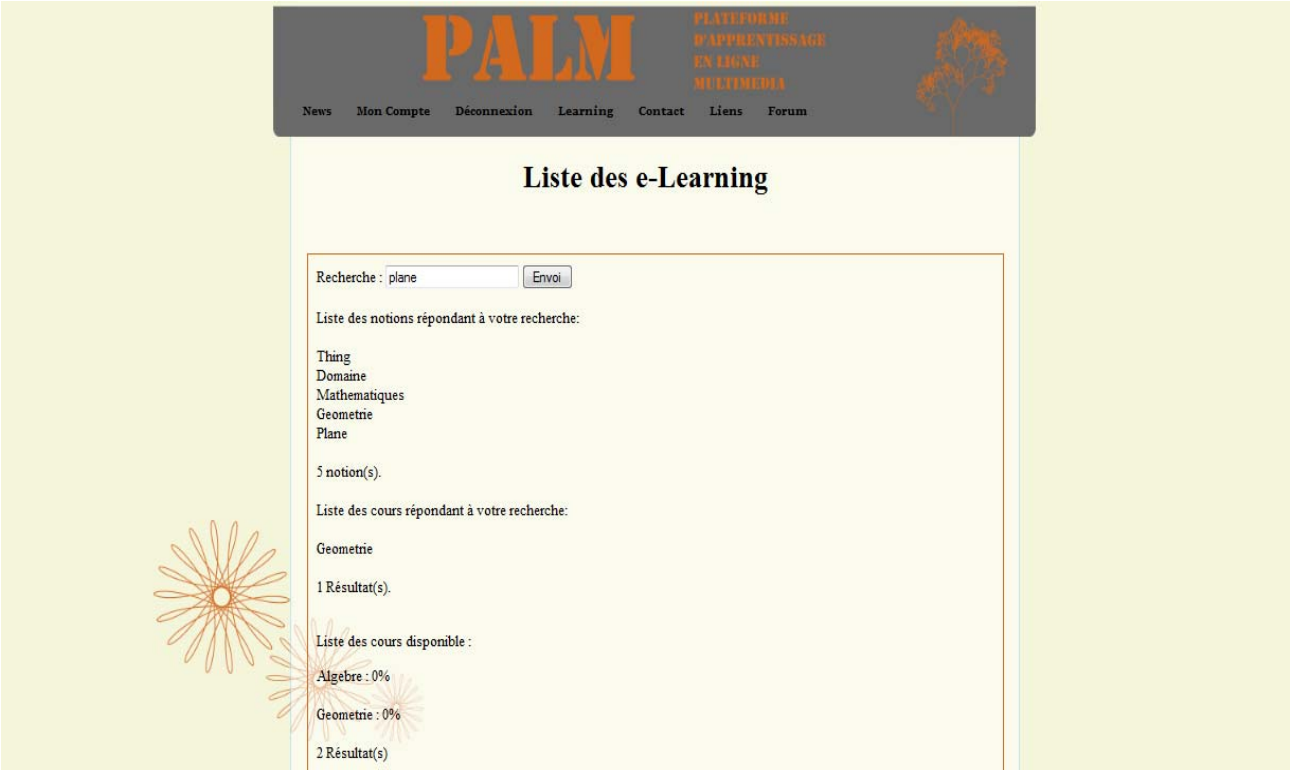

Figure 31 : Interface de recherche de cours.

Ici, l'utilisateur a saisi le terme ***plane*** comme requête. Le système vérifie dans l'ontologie le concept associé au terme plane, puis affiche tous les concepts parents de Plane, qui sont Géométrie, Mathématique. Ensuite, PALM recherche dans sa base les différents cours qui sont indexés par ces concepts. Nous appliquons donc à tout moment une généralisation de la requête utilisateur.

4.1.6. Suivi des connaissances des apprenants

La gestion des connaissances des utilisateurs est l'un des objectifs de notre système d'apprentissage en ligne. L'idée est de mettre en place un système qui mémorise à tout moment l'évolution des connaissances vues par chaque utilisateur.

A chaque consultation de documents du corpus effectué par un utilisateur, matérialisé par la présence des requêtes indexées de l'utilisateur, le système met à jour l'ontologie personnelle de l'utilisateur en fonction des annotations sémantiques des documents qu'il a ouverts au cours de son activité d'apprentissage.

L'utilisateur a la possibilité de valider les nouvelles connaissances acquises après la lecture des documents à l'aide d'une interface de validation de connaissance.

4.1.6.1. Diagramme des cas d'utilisation

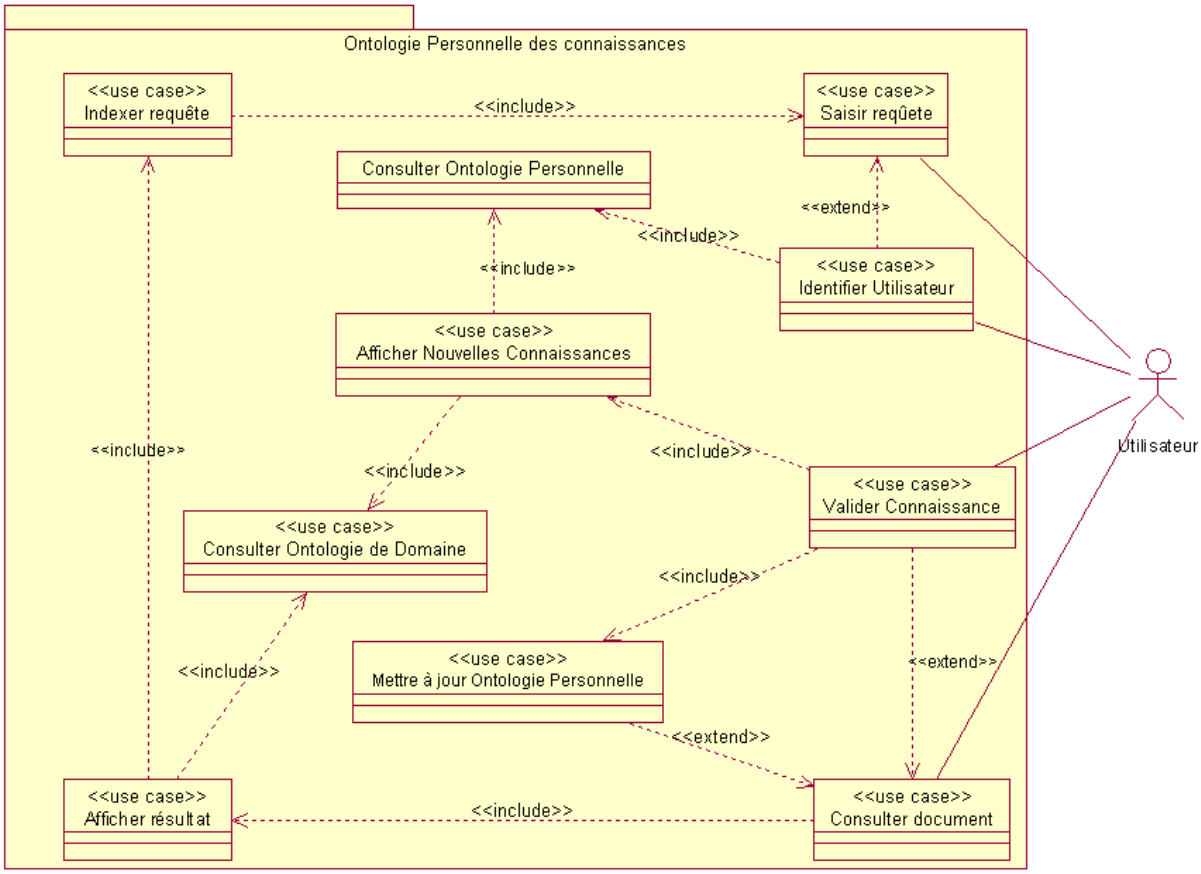

Figure 32. Diagramme des cas d'utilisations du système Ontologie personnelle des connaissances.

Ce diagramme montre l'interaction entre le système « Ontologie Personnelle des Connaissances » et l'utilisateur qui est l'acteur principal.

Comme nous le voyons sur le diagramme, avant de saisir une requête, l'utilisateur doit s'identifier ou doit être identifié. Cela permet au système non seulement de consulter son ontologie personnelle mais aussi de la mettre à jour si l'utilisateur valide les nouvelles connaissances proposées par le système. Ces nouvelles connaissances sont proposées après comparaison de l'ontologie personnelle avec l'ontologie du domaine et les annotations sémantiques dans les documents consultés par l'utilisateur.

4.1.6.2. Suivi des cours

L'utilisateur peut accéder aux cours qui sont disponibles via l'option de menu *Learning*.

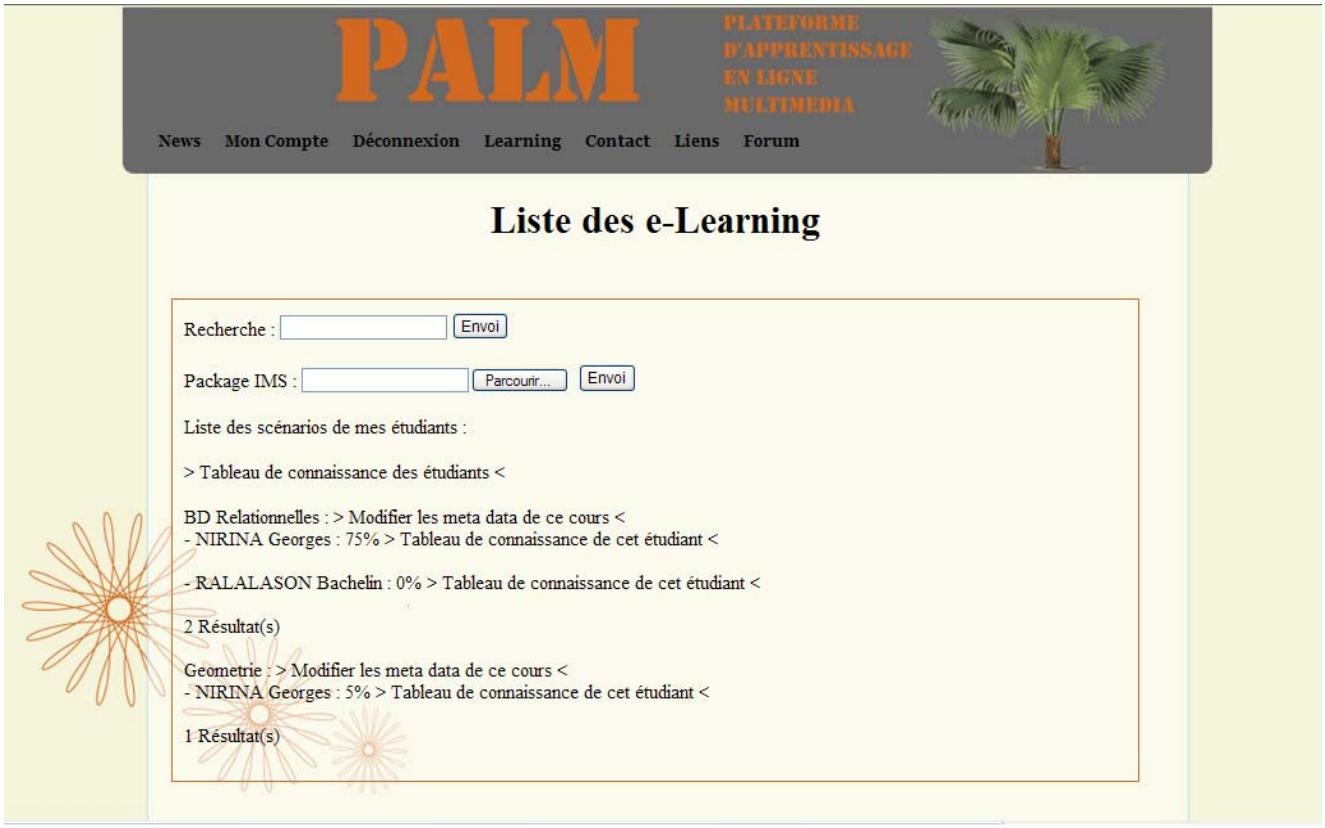

Figure 33 : Accès aux cours.

A côté du nom de chaque cours, un pourcentage s'affiche. Celui-ci représente l'état d'avancement de l'utilisateur sur ce cours. Il peut donc cliquer sur le nom de chaque cours soit pour le commencer soit le continuer.

Par exemple sur cette capture d'écran, nous pouvons constater que l'étudiant NIRINA Georges a déjà fait 75% du cours de BD Relationnelles et 5% du cours de Géométrie, tandis que l'étudiant Bachelin n'a pas encore commencé le cours de BD Relationnelles (0%) et n'est pas inscrit au cours de Géométrie.

A côté du nom de chaque étudiant figure un lien « Tableau de connaissance de cet étudiant » qui permet d'afficher les différentes connaissances acquises par cet étudiant.

4.1.6.3. Suivi des connaissances

Comme indiqué précédemment, au fur et à mesure de l'avancement de l'étudiant dans les activités d'apprentissage, son profil ou ontologie personnelle se forme à travers les notions associées aux objets pédagogiques des activités. Ces profils ou connaissances acquises peuvent être consultés à l'aide du « Tableau de connaissance de cet étudiant » comme l'indique la Figure 34.

Figure 34 : Affichage des connaissances des apprenants.

4.1.7. Conclusion

PALM est une plateforme d'apprentissage en ligne multimédia qui a permis de mettre en œuvre les différentes exigences des systèmes d'apprentissage en ligne dont :

- le suivi et le séquencement des scénarios d'apprentissage,
- l'application des pédagogies d'apprentissages,
- l'annotation sémantique des objets pédagogiques par des concepts issus d'une ontologie de référence,
- la recherche et réutilisation des objets pédagogiques par les enseignants,
- le suivi des connaissances des apprenants,
- l'interaction synchrone et asynchrone entre apprenant et enseignant.

4.2. Le projet Dynamo

L'objectif du projet DYNAMO (DYNAMic Ontology for information retrieval) est de mettre en place un système de recherche sémantique d'information basé sur une ontologie de domaine. L'ontologie est créée à partir des termes venant des textes des documents d'un corpus d'un domaine donné et les concepts de l'ontologie sont utilisés pour annoter et indexer les documents du corpus du domaine.

Contrairement à l'approche classique qui indexe les documents à partir des termes d'un thésaurus ou d'une ontologie prédéfinie ou même sans consultation de vocabulaire de référence, la particularité du projet DYNAMO est l'aspect dynamique des documents du corpus et de l'ontologie. A chaque arrivée de nouveaux documents, le système vérifie si le document pourrait être indexé à partir des concepts de l'ontologie actuelle. Dans le cas contraire, l'ontologie est mise à jour à partir des termes du nouveau document. La nouvelle version de l'ontologie est utilisée pour annoter le nouveau document. Ainsi, les annotations des anciens documents du corpus peuvent être remises en causes ; ces derniers seront alors réindexés.

4.2.1. Les documents à annoter et rechercher

Afin d'assurer la généricité des modules de Dynamo, le projet est expérimenté sur trois domaines d'applications distincts, qui nécessitent la gestion des connaissances évolutives ainsi que la recherche sémantique d'information.

Les domaines d'application qui sont expérimentés dans le projet DYNAMO (projet ANR), en collaboration avec les partenaires qui apportent des données industrielles, sont :

- la RI scientifique en archéologie des techniques, en partenariat avec le laboratoire « Préhistoire et Technologie »,
- le diagnostic en électronique automobile, en partenariat avec la société « ACTIA »,
- la gestion des connaissances associées aux projets informatiques, en partenariat avec l'entreprise « ARTAL Technologies ».

Les partenaires académiques s'intéressent quant à eux à plusieurs aspects relatifs à la construction et mise à jour automatique d'ontologies et à l'indexation et la recherche sémantique. Ce dernier aspect correspond à notre cadre de travail.

Dans ce qui suit, nous illustrons nos propos au travers des exemples issus du partenaire ACTIA dans la mesure où ce sont leurs données qui ont été les premières disponibles. Cependant, nos propositions permettent de prendre en compte les différents cadres applicatifs cités ci-dessus.

4.2.2. Annotation des documents

L'annotation vise à permettre un lien entre le document et les instances des concepts désignés par les occurrences de termes dans les documents. L'annotation d'un document utilise la RTO (Ressource Termino-Ontologique) courante.

Le scénario d'annotation de documents comprend trois alternatives :

- un nouveau document (jamais annoté) doit être annoté avec l'ontologie courante (cas nominal),
- un document déjà annoté avec une version précédente de l'ontologie doit voir son annotation mise à jour (alternative 1),
- un document déjà annoté avec la RTO qui a subi une modification de son contenu doit être ré-annoté.

Le cas d'un document modifié et déjà annoté mais par une version précédente de la RTO est pour le moment traité comme la prise en compte successive des deux scénarios alternatifs. Cependant, dans ce cas précis, il faudra vérifier que les annotations manuelles sont cohérentes avec la nouvelle RTO.

S'il s'agit d'une ré-annotation automatique d'un document, suite à une évolution de la RTO, et lorsque le système demande à l'utilisateur de valider l'annotation, l'ancienne version de l'annotation est affichée afin de l'aider à valider la nouvelle annotation. Il en est de même pour une ré-annotation suite à une modification du document.

4.2.3. Recherche des documents

Pour le cas d'ACTIA, l'analyse sémantique des documents ne porte pas sur tout le texte des documents mais seulement sur une partie, le champ symptôme (cf page 53, section 3.1.3.1). Comparer deux documents (ou un document avec une requête) revient donc à comparer les parties symptômes de ces documents (ou la partie symptôme du document avec la requête). La partie symptôme d'un document peut évoquer un ou plusieurs symptômes. Une requête peut être soit du texte libre saisi par l'utilisateur, soit un document dans la collection. La première étape de la comparaison de documents consiste à extraire les annotations sémantiques dans la partie symptôme du document.

Pour ACTIA, un symptôme est défini comme un ensemble de quatre types de concepts: un problème, une prestation, une relation entre problème et prestation, et les contextes dans lesquels se présente la relation.

Après avoir extrait les différents symptômes des deux documents, l'étape suivante consiste à calculer la similarité sémantique entre les deux documents. La similarité entre deux symptômes est donc considérée comme la combinaison des similarités entre les concepts de même type des deux symptômes.

Le contexte d'apparition d'un symptôme étant un facteur très déterminant dans le diagnostic automobile, nous accordons une grande importance à ce type de concept.

Ainsi, nous pouvons formuler la mesure de similarité entre deux documents D1 et D2 où l'un de ces documents est une requête. Suivant le contexte d'application de la similarité sémantique entre documents, l'une des formules suivantes peut être appliquée.

$$Sim_{Doc}(D_1, D_2) = \sum_{i,j} Sim_{Symp}(S_i, S_j) \quad \textbf{(42)}$$

$$\text{Ou } Sim_{Doc}(D_1, D_2) = \underset{i,j}{Max}\left(Sim_{Symp}(S_i, S_j)\right) \quad \textbf{(43)}$$

$$\text{Ou } Sim_{Doc}(D_1, D_2) = \prod_{i,j} \left(Sim_{Symp}(S_i, S_j) \right) \quad \textbf{(44)}$$

Avec,

$$Sim(S_i, S_j) = \frac{\sum_{i=0}^{taille(S_i)} Coef[n] * Similarité Concept(S_i[n], S_j[n])}{\sum_{n=0}^{taille(S_i)} Coef[n]} \quad \textbf{(45)}$$

Où S_i et S_j sont des symptômes composés de Problème ($S_i[3]$, $S_j[3]$), Prestation ($S_i[2]$, $S_j[2]$) et de Contexte ($S_i[1]$, $S_j[1]$). Chacun de ces types de concept a respectivement son coefficient Coef[n] correspondant. Nous donnons plus d'importance au problème par rapport à la prestation. De même, la prestation est plus importante que le contexte.

La RI commence par l'expression du besoin du lecteur sous forme d'une requête en langage libre. Le lecteur peut également sélectionner un document existant afin de rechercher dans la collection un document sémantiquement similaire à ce dernier.

Le système effectue l'annotation de la requête en mettant en correspondance les occurrences des termes de la requête avec les instances des concepts issus de la RTO. L'annotation d'une requête est équivalente à l'annotation de documents.

Le système recherche alors les documents potentiellement pertinents par rapport à la requête avant de les afficher à l'utilisateur ; celui-ci sélectionne les documents qu'il souhaite voir afficher.

La reformulation de la requête peut intervenir à deux niveaux :

- si lors de la recherche initiale aucun document n'est retrouvé, le système demande la reformulation,
- après une première recherche, si le lecteur n'est pas satisfait de la réponse.

4.2.4. Dynamique de l'ontologie

Comme dans le cas de l'approche sac de mots, la collection de documents à indexer peut être dynamique et donc subir des modifications ; il est donc important de proposer des principes pour la mise à jour des index dans le cas d'une indexation sémantique.

De plus, contrairement à l'approche sac de mots, le vocabulaire utilisé lors de l'indexation peut être amené à varier indépendamment des documents. Ainsi, l'ontologie qui sert de référence à l'indexation peut être modifiée. Dans ce dernier cas, il est important de considérer la mise à jour de l'indexation consécutive à une modification du vocabulaire de référence, cela afin de maintenir une cohérence entre les documents et le vocabulaire d'indexation.

4.2.5. Mise à jour des documents

La mise à jour des documents se fait à l'extérieur du système. Cependant, une fois la nouvelle version du document présente dans le système, une ré-annotation du document est nécessaire.

Plutôt que de ré-annoter l'ensemble du document, nous proposons que seules les parties modifiées le soient. Pour cela, tous les documents sont découpés en blocs pour permettre la gestion des modifications par partie des documents (cf. section 3.2.2, page 64).

4.2.6. Enrichissement du corpus

Afin de pouvoir faire de la recherche, tous les documents du corpus doivent préalablement être annotés. Ainsi chaque document qui va être ajouté au corpus doit être annoté automatiquement. Le système permet également à l'utilisateur de modifier l'annotation manuellement.

Les documents déjà annotés mais qui sont modifiés peuvent subir une ré-annotation. Dans le cas d'un document modifié, les annotations validées manuellement seront affichées.

Les documents déjà annotés doivent être ré-annotés suite à une modification de la RTO afin de garder la cohérence entre les index et la RTO. En effet, non seulement de nouveaux concepts ou relations peuvent apparaître dans la nouvelle version de la RTO, mais également, des concepts ou relations de l'ancienne version peuvent disparaître.

Ci-après le diagramme des cas d'utilisation de l'enrichissement du corpus.

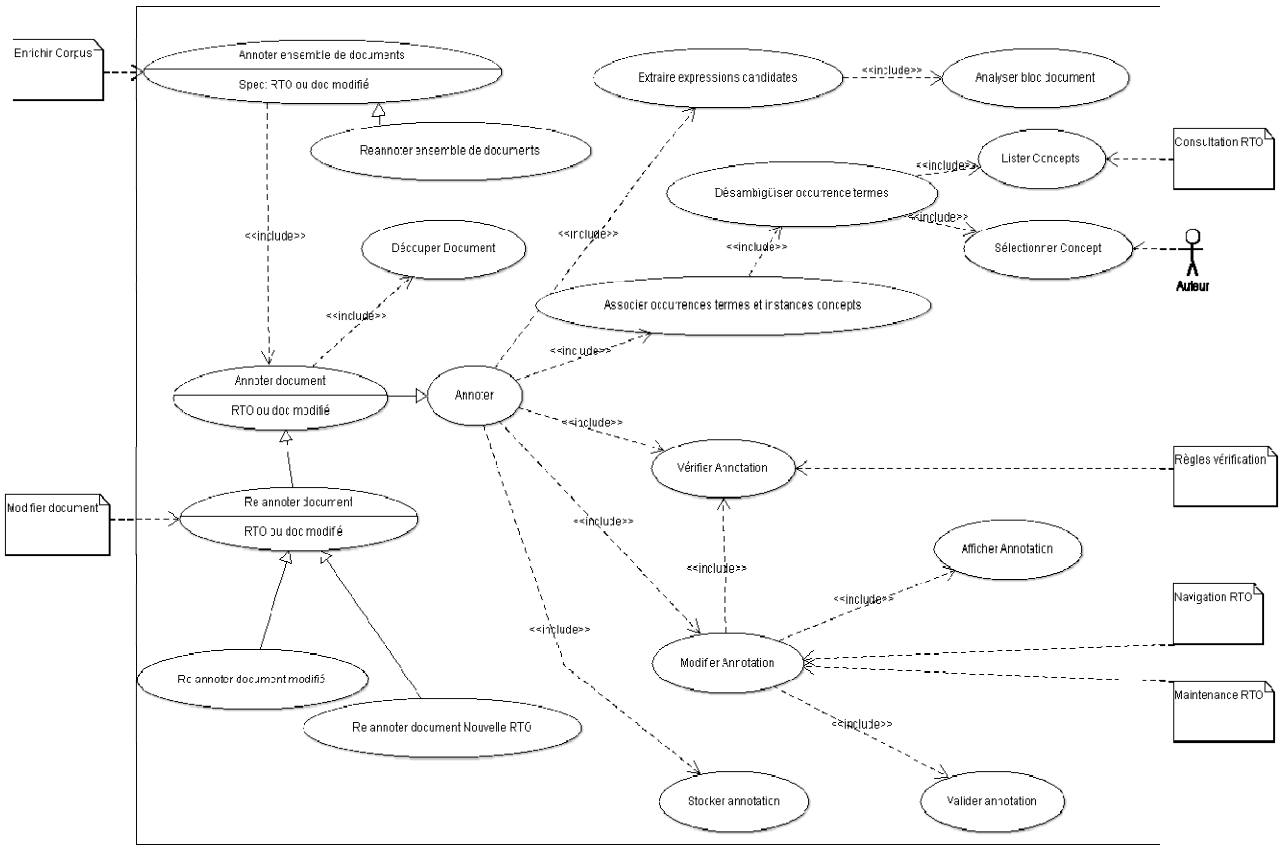

Figure 35 : diagramme des cas d'utilisation de l'enrichissement du corpus.

4.2.7. Expérimentation des mesures de similarité

4.2.7.1. Cadre expérimental

Les mesures de similarité ProxiGénéa que nous avons développées (cf. section 3.4.3) ont été expérimentées pour la recherche des documents dans un corpus de 356 fiches de maintenance automobiles, de la société ACTIA. D'après les résultats de l'évaluation des requêtes que nous détaillons ci-après, nous pouvons conclure que la mesure de similarité conceptuelle (ProxiGénéa) associée avec la fonction de similarité sémantique entre requête et document (GenericSimilarity) que nous avons proposée permet de retourner tous les documents pertinents. GenericSimilarity, comme son nom l'indique, est une fonction de similarité générique qui peut être appliquée sur toute ontologie et sur toute collection de documents annotées avec les concepts de l'ontologie en question. (Ceci nous a permis de l'appliquer dans d'autres cadres applicatifs des partenaires du projet dont ARTAL et Préhistoire et Technologie).

Nous pouvons aussi constater, la diminution très considérable du bruit documentaire.

Nous détaillons ci-après les résultats de l'expérimentation de nos mesures de similarité dans le cadre du projet Dynamo.

Voici les vingt requêtes exécutées sur le SRI Dynamo avec le corpus d'ACTIA.

1. Calage moteur en décélération au ralenti.
2. Manque de puissance moteur avec allumage du voyant diagnostic.
3. Démarrage difficile à froid.
4. Témoin diag. allumé constamment.
5. Témoin diag. allumé en permanence avec message sur EMF: anomalie anti pollution.
6. La batterie se décharge en quelques heures.
7. La ventilation du chauffage est hors service.
8. Le témoin d'airbag s'allume par intermittence.
9. Le moteur fait des à-coups, cale au ralenti et le témoin d'injection s'allume.
10. Le moteur fonctionne avec un mélange très riche par intermittence.
11. Le moteur boite légèrement, le ralenti est instable.
12. Pas d'allumage du témoin d'injection à la mise du contact.
13. Le moteur fait des ratés en début d'accélération sans allumage du témoin défaut.
14. Le moteur ne démarre pas de façon intermittente.
15. Le moteur surchauffe sur autoroute, surtout en forte charge.
16. Le moteur tourne avec des trous en début de reprise.
17. Le moteur s'étouffe par intermittence.
18. Emission de fumée noire et manque de puissance.
19. La climatisation ne fonctionne plus.
20. Le moteur tombe en panne en roulant à chaud.

4.2.7.2. Mesures

Nous présentons dans cette sous section les mesures de performance des SRI que nous utilisons pour évaluer les mesures de similarité que nous avons proposées.

Les mesures Rappel / Précision :

Le rappel et la précision sont mesurés après que le système détermine un classement sur les documents de la collection en réponse à la requête de l'utilisateur. Ils sont obtenus en partitionnant l'ensemble des documents restitués par le SRI en deux catégories appelés « les documents pertinents » et « les documents non pertinents ». Ces deux mesures sont définies comme suit :

- Précision : la précision mesure la capacité du système à rejeter les documents non pertinents par rapport à une requête. Il est donné par le rapport entre l'ensemble des documents pertinents retrouvés et l'ensemble des documents sélectionnés.
- Rappel : le rappel mesure la capacité du système à retrouver tous les documents pertinents par rapport à une requête. Le rappel est défini par le rapport entre le nombre de documents pertinents retrouvés au regard du nombre de documents pertinents que possède la collection (Salton, 1971).

$$Rappel = \frac{a}{a+c} \qquad \Pr écision = \frac{a}{a+b}$$

Où :

- *a* est le nombre de documents pertinents retrouvés,
- *b* est le nombre de documents non pertinents retrouvés,
- *c* est le nombre de documents pertinents non retrouvés,
- *d* est nombre de documents non pertinents non retrouvés.

La courbe Rappel / Précision

La courbe rappel / précision est souvent utilisée comme une mesure combinée d'évaluation des SRI. Une telle courbe donne la valeur de précision correspondant à chaque point de rappel.

Précision à x documents

Il s'agit de la valeur de la précision obtenue pour les x premiers documents de la liste ordonnée des documents restitués.

Précision Moyenne Globale

La précision moyenne est la moyenne des valeurs de précision calculée à chaque document pertinent de la liste ordonnée. Si un document pertinent est retourné à la $x^{ème}$ position, la précision pour ce document est « la précision à x documents ». Si un document pertinent n'a pas été trouvé par le système, la précision pour ce document est nulle. La MAP se calcule comme suit :

$$MAP = \sum_{i=1}^{N_q} \frac{P_i(X)}{N_q}$$

Où :

- N_q est le nombre total de requêtes,
- $P_i(X)$ donne la précision de la i[ème] requête correspondant au taux de rappel *X*.

4.2.7.3. Résultats

Nous avons choisi d'appliquer sur la fonction de similarité sémantique GenericSimilarity les différentes mesures de similarité conceptuelle afin d'évaluer leurs performances en termes de rappel-précision.

Rappel	Classique	GenericSimilarity (Wu et Palmer)	GenericSimilarity (ProxiGénéa)	GenericSimilarity (ProxiGénéa2)	GenericSimilarity (ProxiGénéa3)
0	0,7891	0,9437	0,9625	0,9625	0,9750
0,1	0,7678	0,9199	0,9487	0,9477	0,9602
0,2	0,6220	0,8721	0,8998	0,8986	0,9136
0,3	0,5530	0,8213	0,8401	0,8472	0,8625
0,4	0,5269	0,7945	0,8401	0,8247	0,8320
0,5	0,4372	0,7700	0,8192	0,8141	0,8305
0,6	0,4181	0,7074	0,7474	0,7651	0,7856
0,7	0,3754	0,6764	0,7283	0,7511	0,7404
0,8	0,3239	0,6400	0,7026	0,7224	0,7279
0,9	0,2951	0,6023	0,6642	0,6587	0,6833
1	0,2643	0,5471	0,6099	0,6092	0,6138

Tableau 3 : Comparaison des valeurs des rappels / précision par fonction de similarité

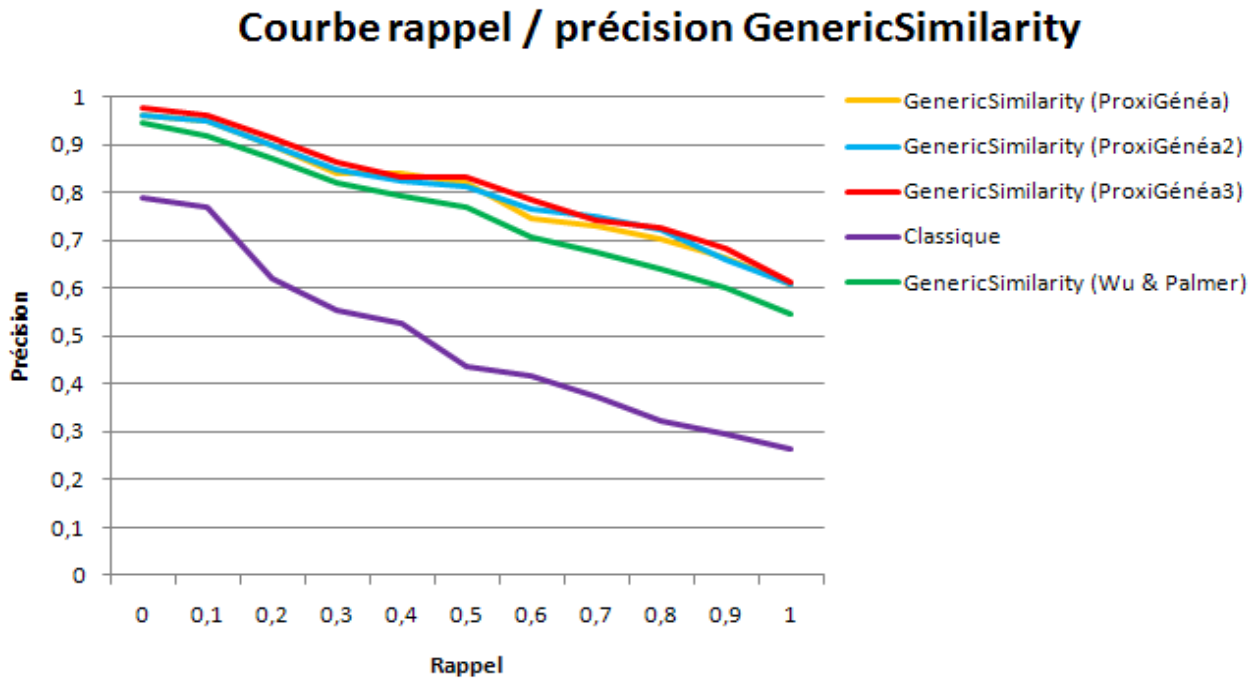

Figure 36 : Comparaison des valeurs des Rappels / précision par fonction de similarité

L'analyse de ces courbes montre la supériorité des trois variantes de ProxiGénéa par rapport à (Wu et Palmer, 1994) en matière de rappel et précision. Nous constatons aussi la nette prédominance de la RI sémantique par rapport à la RI classique.

Le tableau 4 présente la précision moyenne, toutes requêtes confondues, pour chaque mesure de similarité conceptuelle.

	Toutes les requêtes
Classique	0,4484
GenericSimilarity (ProxiGénéa)	0,7718
GenericSimilarity (Wu et Palmer)	0,7261
GenericSimilarity (ProxiGénéa2)	0,7741
GenericSimilarity (ProxiGénéa3)	0,7890

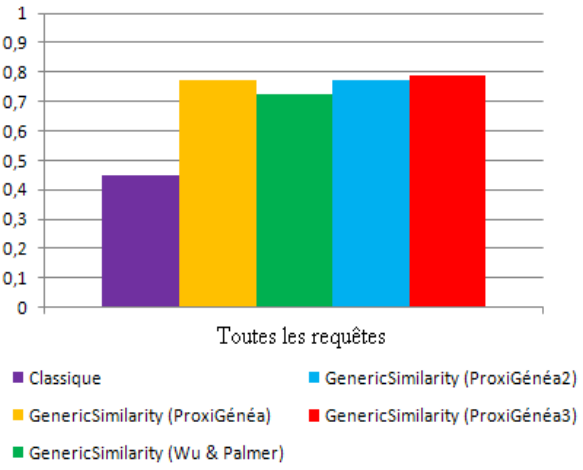

Tableau 4 : Comparaison des Précisions moyenne globale

Ainsi, la précision moyenne globale, qui donne la moyenne des valeurs de précision à chaque document pertinent de la liste ordonnée, montre que ProxiGénéa3 est la plus précise de ces mesures de similarité. D'ailleurs, les variantes de ProxiGénéa restent toutes les trois au-dessus de (Wu et Palmer, 1994).

Le tableau 5 indique la précision moyenne pour chaque requête et pour chaque mesure de similarité.

	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10
Classique	0,25	0,47	0,80	0,43	0,68	0,97	0,88	0,22	0,60	0,24
GenericSimilarity (ProxiGénéa)	0,75	0,99	0,84	0,61	0,89	1,00	1,00	0,98	0,95	0,29
GenericSimilarity (Wu et Palmer)	0,75	0,99	0,83	0,51	0,89	1,00	1,00	0,98	0,92	0,30
GenericSimilarity (ProxiGénéa2)	0,75	0,99	0,84	0,53	0,89	1,00	1,00	0,98	0,95	0,29
GenericSimilarity (ProxiGénéa3)	0,75	0,99	0,79	0,53	0,89	1,00	1,00	0,98	0,87	0,29

	R11	R12	R13	R14	R15	R16	R17	R18	R19	R20
Classique	0,60	0,02	0,11	0,26	0,36	0,64	0,20	0,43	0,46	0,36
GenericSimilarity (ProxiGénéa)	0,62	0,42	0,26	0,92	0,64	0,81	0,84	1,00	0,69	0,95
GenericSimilarity (Wu et Palmer)	0,57	0,42	0,27	0,92	0,28	0,80	0,56	1,00	0,69	0,85
GenericSimilarity (ProxiGénéa2)	0,61	0,42	0,25	0,92	0,64	0,95	0,85	1,00	0,69	0,95
GenericSimilarity (ProxiGénéa3)	0,59	0,42	0,28	0,92	1,00	0,95	0,89	1,00	0,69	0,97

Tableau 5 : Valeur des précisions moyenne par requête.

Pour faciliter la lecture, les plus grandes valeurs en MAP pour chaque requête ont été mises en relief. Parmi les vingt requêtes soumises au système Dynamo, Wu et Palmer dépasse les fonctions ProxiGénéa uniquement pour la requête R10. Pour cette requête, la valeur des précisions pour chaque mesure de similarité ne dépasse pas 0,30.

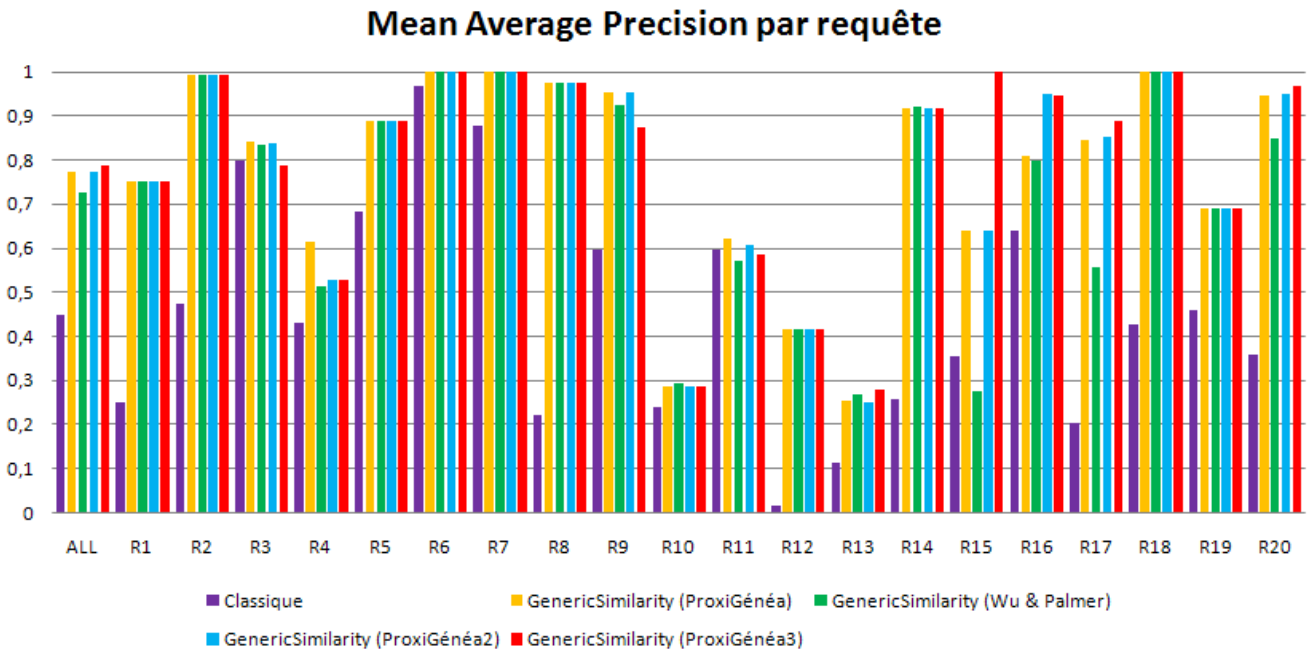

Figure 37 : Précision moyenne par requête.

Le tableau 6 montre les précisions à x documents pour chaque mesure de similarité conceptuelle appliquée dans la fonction de similarité sémantique GenericSimilarity.

Précision	Classique	GenericSimilarity (ProxiGénéa)	GenericSimilarity (ProxiGénéa2)	GenericSimilarity (ProxiGénéa3)	GenericSimilarity (Wu et Palmer)
P1	0,6500	0,9000	0,9000	0,9500	0,9000
P2	0,5750	0,8750	0,8750	0,9000	0,8250
P5	0,5000	0,7700	0,7800	0,7800	0,7300
P10	0,3650	0,5700	0,5750	0,5700	0,5550
P15	0,2900	0,4700	0,4667	0,4633	0,4267
P20	0,2525	0,4025	0,3975	0,3975	0,3700
P30	0,2033	0,3083	0,3017	0,3100	0,2883
P50	0,1690	0,2080	0,2080	0,2080	0,2050

Tableau 6 : Valeur des précisions à x documents par fonction de similarité sur GenericSimilarity.

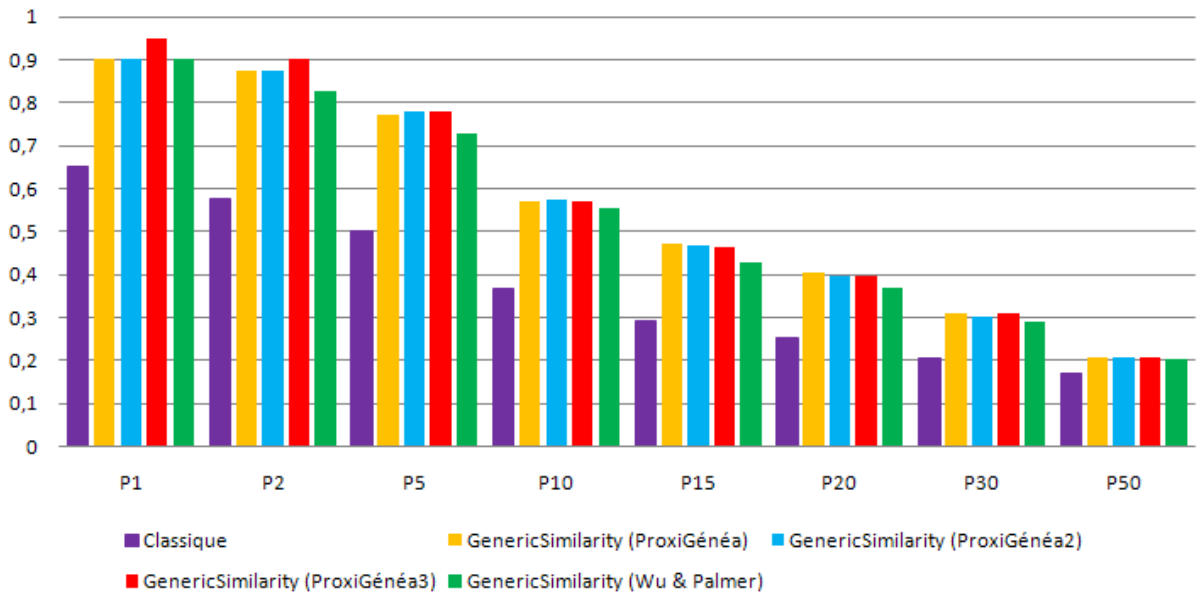

Figure 38 : précisions à x documents par fonction de similarité sur GenericSimilarity.

Globalement, nous constatons que pour la haute précision, ProxiGénéa3 est la plus efficace par rapport aux autres mesures de similarité conceptuelle. Cependant, à partir de la précision à 10 documents (P10), ProxiGénéa prend le dessus. Ces caractéristiques sont intéressantes pour pouvoir choisir laquelle des mesures de similarité conceptuelle doit être choisie selon les niveaux de précision voulus.

Nous avons également appliqué les différentes mesures de similarité conceptuelle dans la fonction de similarité propriétaire d'ACTIA (OwnSimilarity) qui implémente initialement la mesure de similarité Wu et Palmer. La particularité de cette fonction est qu'elle est totalement conçue et adaptée à l'ontologie de maintenance automobile d'Actia. Elle n'est pas applicable à une autre ontologie.

Rappel	Classique	OwnSimilarity (Wu et Palmer)	OwnSimilarity (PoxiGénéa)	OwnSimilarity (PoxiGénéa2)	OwnSimilarity (PoxiGénéa3)
0	0,7891	0,9750	0,9750	0,9750	0,9750
0,1	0,7678	0,9707	0,9707	0,9707	0,9727
0,2	0,6220	0,9266	0,9266	0,9266	0,9287
0,3	0,5530	0,8790	0,8682	0,8696	0,8741
0,4	0,5269	0,8601	0,8575	0,8594	0,8627
0,5	0,4372	0,8353	0,8402	0,8408	0,8388
0,6	0,4181	0,8006	0,8066	0,8065	0,7942
0,7	0,3754	0,7835	0,7572	0,7572	0,7623
0,8	0,3239	0,7384	0,7514	0,7505	0,7523
0,9	0,2951	0,7048	0,7257	0,7072	0,6726
1	0,2643	0,6498	0,6593	0,6401	0,6011

Tableau 7 : Valeur des précisions / Rappel de OwnSimilarity.

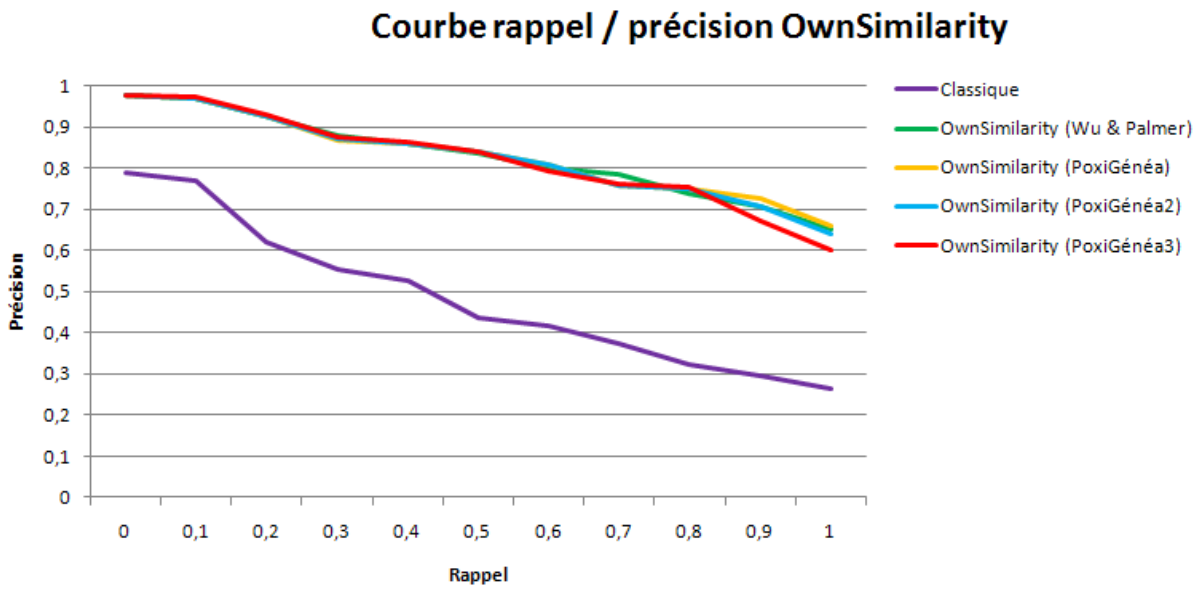

Figure 39 : Courbes Précisions / Rappel de OwnSimilarity

Nous constatons que les quatre mesures de similarité conceptuelle ont à peu près le même comportement. Cependant, ProxiGénéa3 reste au dessus des autres pour les hautes précisions.

Le tableau 8 monte les précisions moyennes de la fonction de similarité sémantique OwnSimilarity utilisant les différentes mesures de similarité conceptuelle. Nous rappelons que la version actuelle d'OwnSimilarity utilise Wu et Palmer comme mesure de similarité conceptuelle.

	Toutes les requêtes
Classique	0,4484
OwnSimilarity (Wu et Palmer)	0,8143
OwnSimilarity (ProxiGenea)	0,8152
OwnSimilarity (ProxiGenea2)	0,8125
OwnSimilarity (ProxiGenea3)	0,8093

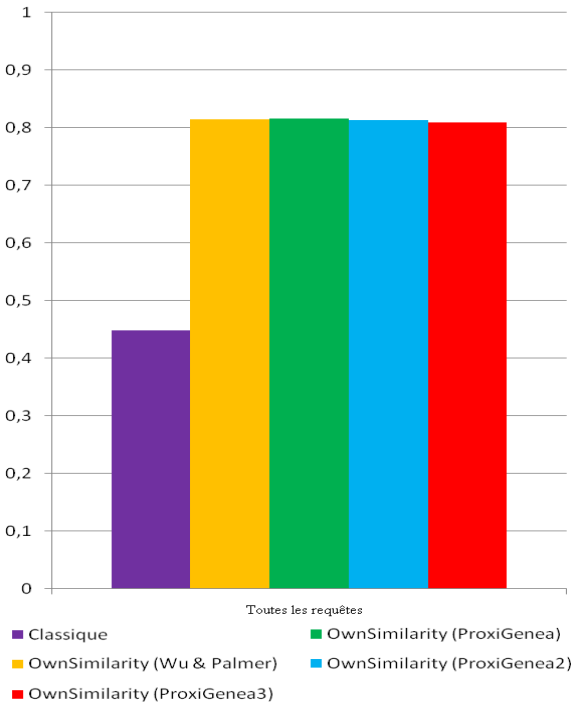

Tableau 8 : Valeur des précisions moyennes de la fonction de similarité OwnSimilarity suivant les mesures de similarités conceptuelles.

Pour toutes les requêtes, la moyenne des précisions montre la supériorité de ProxiGénéa par rapport aux autres mesures de similarité conceptuelle.

Le tableau 9 montre les précisions à x documents pour chaque mesure de similarité conceptuelle appliquée dans la fonction de similarité sémantique propriétaire OwnSimilarity.

	Classique	ProxiGénéa	ProxiGénéa2	ProxiGénéa3	Wu et Palmer
P1	0,6500	0,9500	0,9500	0,9500	0,9500
P2	0,5750	0,9250	0,9250	0,9250	0,9250
P5	0,5000	0,7700	0,7700	0,7800	0,7700
P10	0,3650	0,6200	0,6200	0,6250	0,6200
P15	0,2900	0,5133	0,5067	0,5067	0,5200
P20	0,2525	0,4325	0,4300	0,4275	0,4375
P30	0,2033	0,3217	0,3217	0,3217	0,3150
P50	0,1690	0,2120	0,2130	0,2090	0,2120

Tableau 9 : Valeur des Précisions à x documents par fonction de similarité (OwnSimilarity).

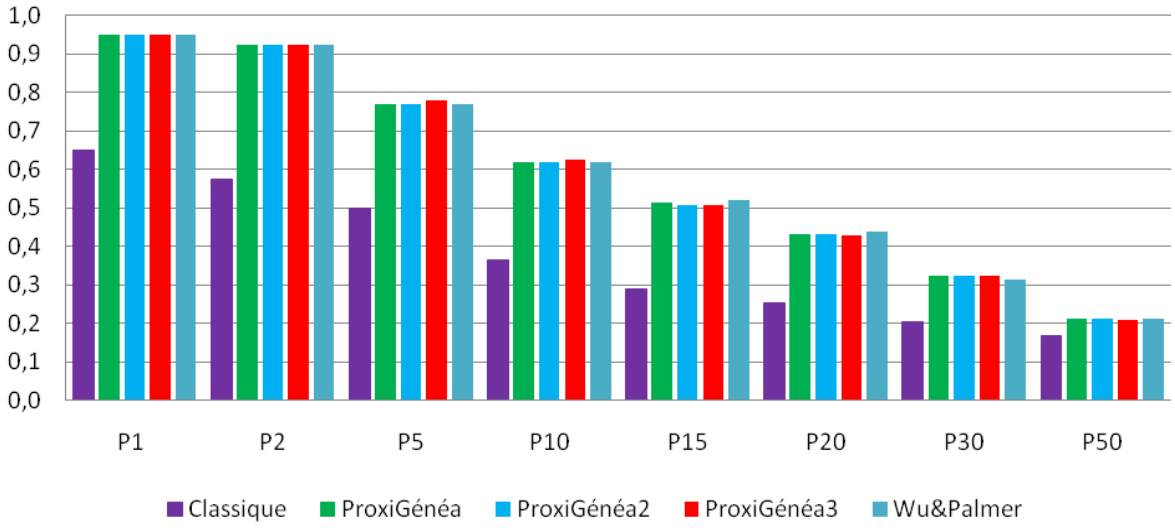

Figure 40 : Précisions à x documents par fonction de similarité (OwnSimilarity)

L'analyse du tableau montre que globalement, ProxiGénéa3 est la plus efficace, pour la haute précision, par rapport aux autres mesures de similarité conceptuelle quand elles sont toutes associées avec la fonction de similarité sémantique OwnSimilarity. Ensuite, ProxiGénéa3 est supérieure à partir de la précision à 30 documents retournés (P30).

4.2.7.4. Conclusion

Le résultat des évaluations de nos mesures de similarité dans le cadre du projet Dynamo montre que ProxiGénéa permet d'obtenir de meilleurs résultats que Wu et palmer en termes de Rappel / Précision. Ceci peut être du au fait que nous avons pris en compte, dans la formulation de nos fonctions de similarité conceptuelle, la sémantique des concepts en tenant compte de la hiérarchie de concepts dans les ontologies.

En outre, nous pouvons aussi constater la nette supériorité de la RI sémantique par rapport à la RI classique dans le cadre de notre application.

4.3. BMML et le Projet Contrapunctus

La compréhension des sens des contenus des documents dépend du format dans lequel on peut percevoir l'information. Ainsi, afin que les malvoyants puissent accéder et comprendre les contenus des partitions musicales, une représentation adaptée de ces documents doit être mise à leur disposition. Le projet Contrapunctus que nous allons décrire ci-après a pour objectif de donner aux malvoyants la possibilité d'accéder et de comprendre les contenus des partitions Braille.

4.3.1. Le projet Contrapunctus

Les progrès réalisés dans le cadre de l'accès à l'information textuelle ne suffisent pas pour permettre à un non voyant d'accéder à une partition musicale.

Contrapunctus est un projet de recherche Européen dont l'objectif principal est de concevoir et développer un service permettant aux musiciens handicapés visuels ou non voyants d'utiliser et d'accéder plus rapidement et plus facilement aux partitions musicales Braille qui se trouvent dans les bibliothèques et les centres de transcriptions. Le travail réalisé dans le cadre du projet *Contrapunctus* débute par une analyse des besoins et focalise son attention sur la réalisation d'un code numérique spécifique pour le Braille musical. Ce travail est motivé par le manque actuel de normalisation des différents supports numériques utilisés dans les bibliothèques Braille. La création d'un code unifié permettra l'échange de partitions musicales et la transformation des partitions existantes qui pourra se réaliser de manière quasiment automatique.

Pour atteindre cet objectif, le projet prévoit la conception et le développement d'outils innovants permettant d'archiver la musique en Braille dans les plus importantes bibliothèques européennes. Ce système permettra également d'échanger les textes à travers Internet et offrira de nouvelles solutions pour lire, comprendre et mémoriser la musique.

Ainsi, il a été décidé de développer :

- un code XML orienté musique Braille,
- un logiciel capable de lire et interpréter les partitions musicales Braille,
- un logiciel flexible pour un utilisateur non voyant lui permettant d'accéder à la musique digitalisée, de lire, manipuler, imprimer des partitions, en fonction de ses besoins spécifiques (débutant, professeur de musique aveugle, amateur …),
- portail d'accès aux bibliothèques de partitions musicales Braille.

Notre participation dans ce projet est essentiellement dans la première partie qui est la spécification, la conception et le développement de ce code musical Braille (BMML).

4.3.2. Partition Musicale Braille

Une partition musicale ne se limite pas à une suite de notes. Il y a également diverses informations importantes qui indiquent la façon dont un musicien doit lire ces notes. Parmi ces informations figurent la clé, l'armure, la signature rythmique, les signes de liaisons, les dynamiques et nuances. Ainsi, la lecture de la musique prend en compte les deux dimensions verticale et horizontale. La difficulté du Braille est que l'on doit transcrire un document à lecture en "deux dimensions" en une suite linéaire de caractères.

4.3.3. Code BMML

Le langage BMML que nous avons proposé (Encelle et al, 2008) permet d'enregistrer et de structurer l'information contenue dans une partition Braille mais également d'en conserver la présentation. Une présentation sous forme graphique devra pouvoir être recréée automatiquement. La production d'une partition sonore au format Midi devra également être possible.

Nous détaillons ici les spécificités de ce code en ce qui concerne l'utilisation de métadonnées non contenues dans les codes existants. La structure ainsi que les contractions spécifiques aux partitions Braille sont ensuite exposées et nous exprimons comment elles sont prises en compte dans ce code.

4.3.3.1. Les métadonnées :

Les informations classiquement contenues dans les métadonnées à savoir le titre, le nom de l'auteur, l'année, l'instrument … sont enrichies d'informations propres au Braille comme par exemple le nombre de caractères par page, le nombre de lignes par page … L'édition de la partition qui a permis de réaliser la transcription est aussi une information utile dans le cas d'un travail entre musiciens voyants et non voyants.

La structure d'une partition Braille est respectée. Pour un instrument donné, une partie est constituée d'un ensemble de mesures reliées par des connecteurs. Une mesure est un ensemble de notes précédées et suivies d'informations les caractérisant comme par exemple le doigté, les liaisons, les nuances ...

Cette structure de la musique peut se retrouver dans les codes classiques de description de la musique comme par exemple dans Musicxml[8]. Il est plus intéressant de caractériser les informations directement liées au Braille comme par exemple les contractions qui se trouvent en Braille pour réduire le nombre de caractères lus ainsi que le nombre de points d'un caractère Braille. Ces abréviations sont réalisées dans le but de faciliter la lecture et la mémorisation d'une partition. Elles sont prises en compte dans le code comme nous allons le voir dans l'exemple suivant.

Une des règles de contraction consiste à doubler une information avant une séquence et à la réécrire une fois à la fin. Par exemple, lorsque plus de 4 accords de même valeur se succèdent, le premier symbole est doublé et repris une dernière fois à la fin. Évidemment cette règle de contraction peut ou non être suivie par le transcripteur mais dans le cas où celle-ci est utilisée nous avons choisi de la coder tel que et c'est le programme de lecture qui prendra en charge la restitution du contenu associé.

Cette restitution est nécessaire pour produire la partition dans une forme classique graphique et également en Midi.

4.3.3.2. Schéma BMML

Une partition Braille est composée des éléments des métadonnées et d'une ou plusieurs parties qui sont subdivisées en sections, lyriques et symboles des accords. En plus de toutes les métadonnées, le langage BMML que nous proposons contient spécifiquement d'autres éléments comme :

- le lien vers une partition en noire (la partition en noire correspondante),

[8] http://www.recordare.com/xml.html

- le lien vers la source (un lien vers une partition Braille source) à partir de laquelle la partition est écrite, la liaison Audio (un lien vers différentes versions de fichiers audio correspondant à la partition),
- ISBDPM (International Standard Bibliographic Description for Printed Music) (ISBDPM) ou la description bibliographique internationale normalisée de la musique imprimée,
- et enfin les informations spécifiques concernant l'affichage de la partition telles que le nombre de lignes dans une page et le nombre de caractères par ligne.

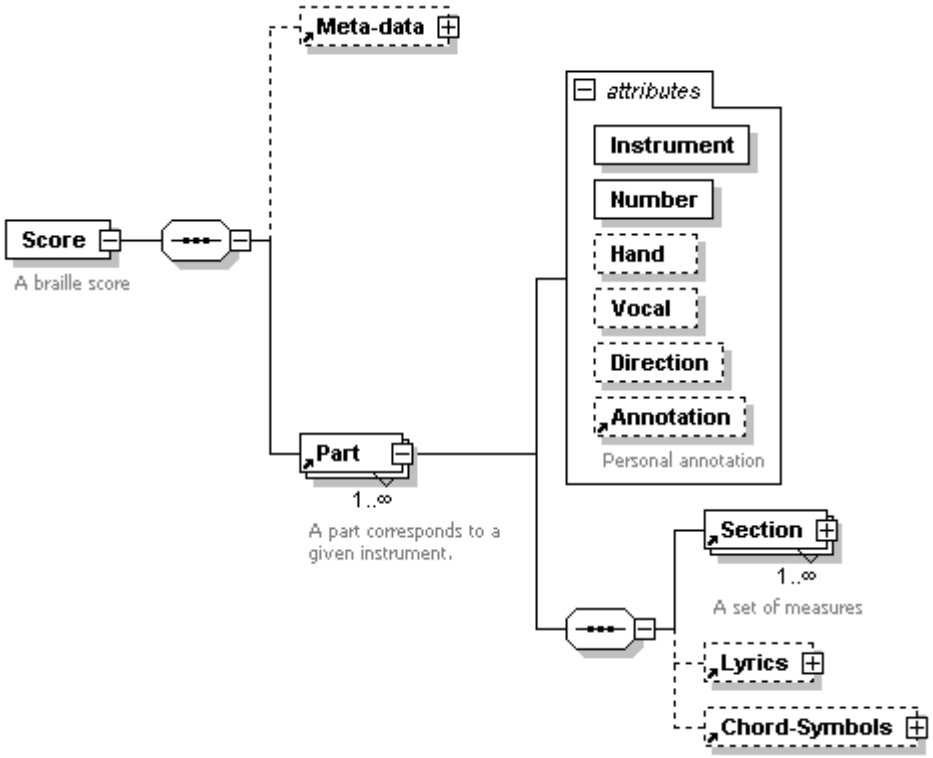

Figure 41 : Schéma du BMML.

4.3.3.3. Métadonnées dans BMML

Les métadonnées utilisées dans le schéma BMML sont décrites dans la Figure 42. Toutes les métadonnées sont contenues dans les documents. Cependant, il sera possible de les exporter dans une base de données pour faciliter l'indexation. Inversement, il sera possible de valoriser les métadonnées à partir d'une base de données.

Le champ *Miscelaneous-field* des métadonnées permet d'insérer de nouvelles métadonnées qui ne sont pas prévues dans la structure initiale de BMML. Chaque utilisateur est donc libre de créer dans sa partition de nouvelles métadonnées propres à son application.

Le champ *Title* indique un nom donné à la ressource. Le champ *Subject* indique le thème de la partition.

Le champ *Creator* qui indique les créateurs de l'œuvre est subdivisé en trois sous-champs dont le champ *Autor* (auteur), *Composer* (compositeur) et *Arranger* (arrangeur). L'élément *Line-In-a-Page* désigne le nombre de lignes par page tandis que *Page-number* indique le nombre total de pages utilisées par la partition.

Le champ *Description* est un texte libre qui décrit le contenu de la partition.

Le champ *Language* décrit le langage dans lequel la partition a été écrite.

Le champ *Display* qui indique le format de la partition est subdivisé en trois sous-champs dont Number-Of-Symbol-In-a-Line (Nombre de caractères Braille sur une ligne), *Number-Of-*

Le champ *Encoding* regroupe les informations relatives à l'encodage de la partition. Parmi ces informations, nous trouvons le champ *Encoding-date* qui indique la date de création de la partition, *Encoder* qui indique le nom de celui qui a édité la partition, *Software* qui indique le nom du logiciel avec lequel la partition a été créée, *Encoding-Description* qui est un texte libre décrivant le processus d'encodage, *Contributor* qui désigne les entités qui ont contribué à la réalisation de l'encodage de l'œuvre et *Support* qui indique le support sur lequel la partition a été encodée.

Le champ *Right* concerne les différents droits relatifs à la partition considérée. Il est composé du champ *License* qui indique le droit de licence sur l'œuvre encodée et du champ *Copyright* qui indique le droit de reproduction réservé sur l'œuvre.

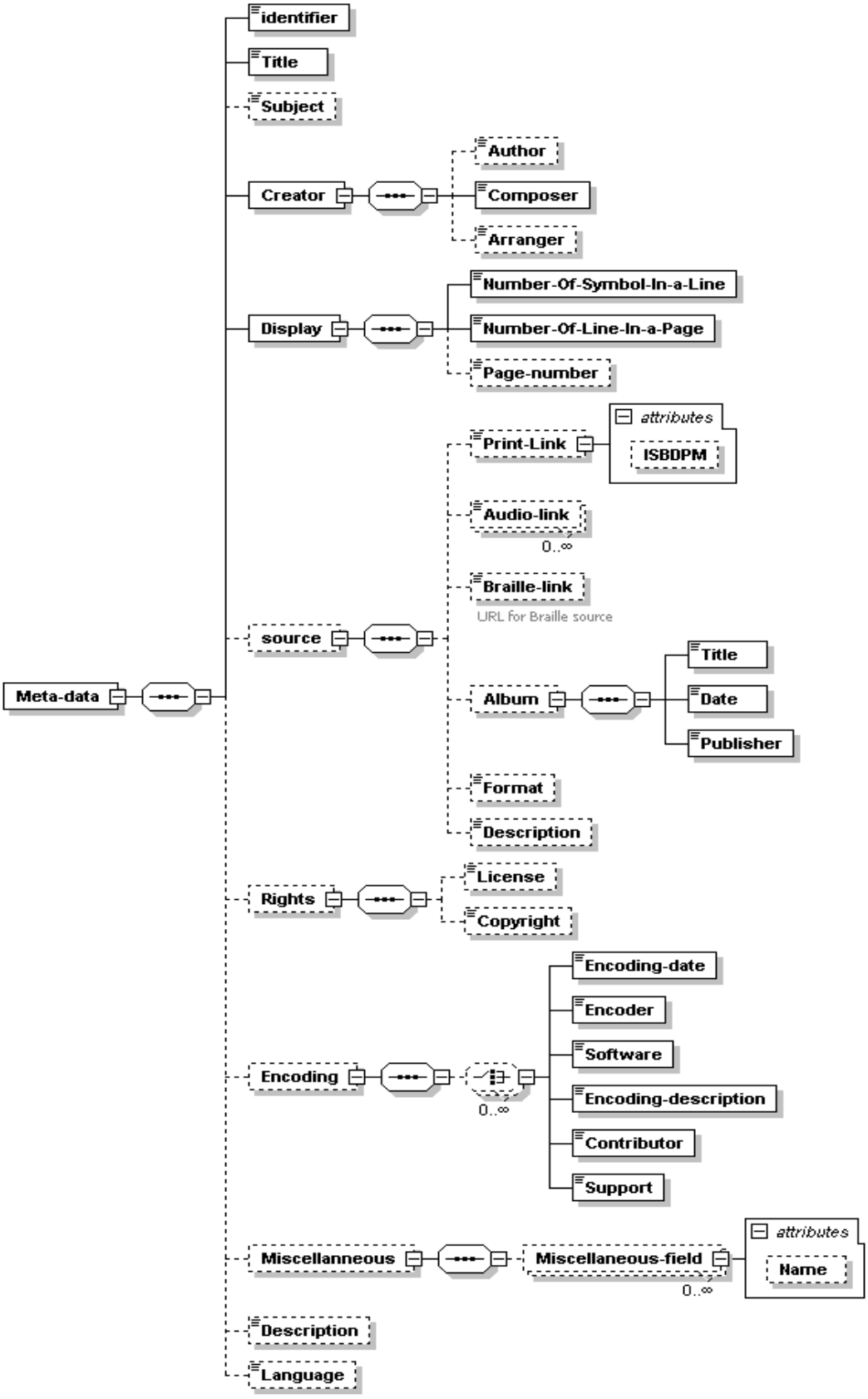

Figure 42 : Les métadonnées dans BMML.

Chaque section est caractérisée par sa clef, l'armure de clef et l'indication de la mesure. Une section est composée d'une ou plusieurs mesures.

Comme une propriété, une section est décrite par son mouvement (grave, largo, larghetto, lento, adagio…), le tempo (80, 100, 120 …), la longueur (ronde/quadruples croche, blancs/triples croches, noirs/double croche …) et son numéro, comme indiqué dans la Figure 43.

Chaque section peut être annotée à l'aide des annotations personnelles. Ces informations peuvent êtres utilisées pour rechercher des partitions musicales ayant certaines caractéristiques.

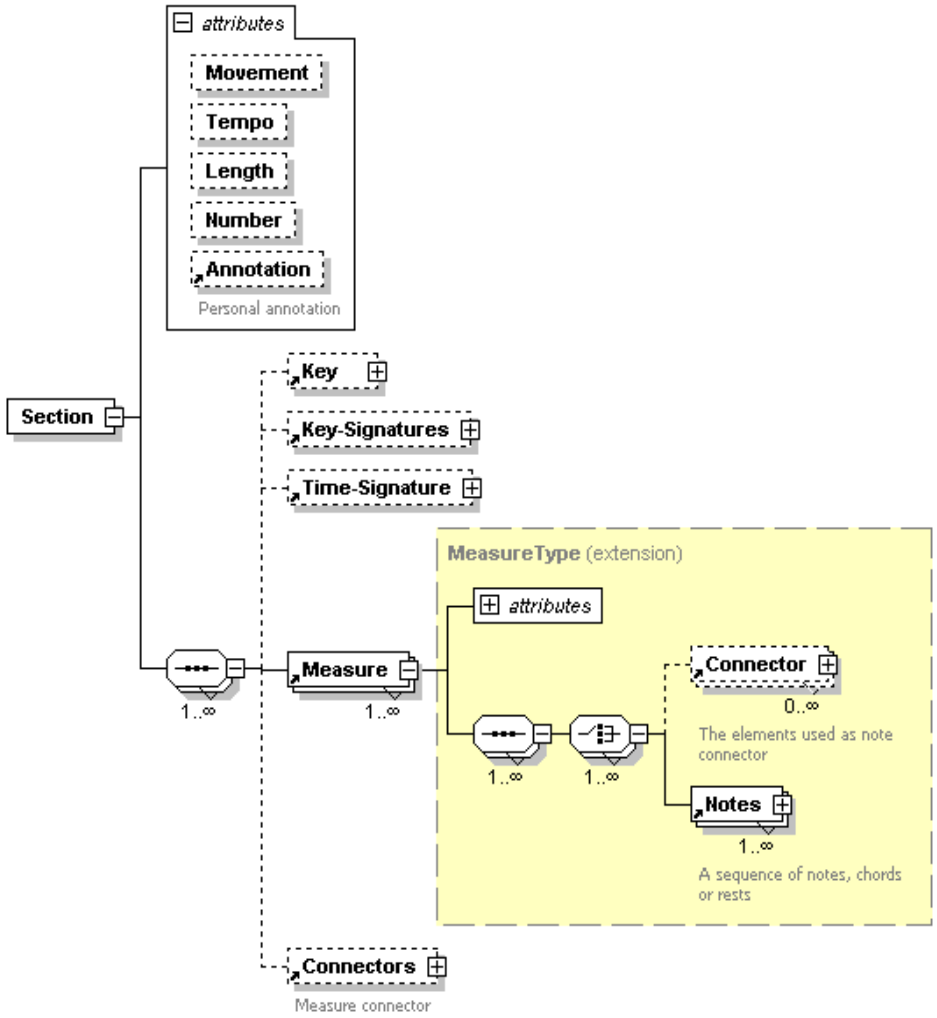

Figure 43 : Structure d'une section dans une partition Braille.

Dans une partition Braille, l'armure de clé (Key signature), qui exprime la tonalité de la chanson, est décrite en spécifiant le nombre d'altérations (dièses et bémols) mais pas la hauteur réelle comme c'est le cas dans une partition en noir. (cf. Figure 44)

La propriété *Concise-Presentation* spécifie si le signe d'altération sera affiché avec une présentation concise ou pas.

La propriété *Number* spécifie le nombre de signes d'altération ; un signe peut être un dièse, un bémol ou un bécarre.

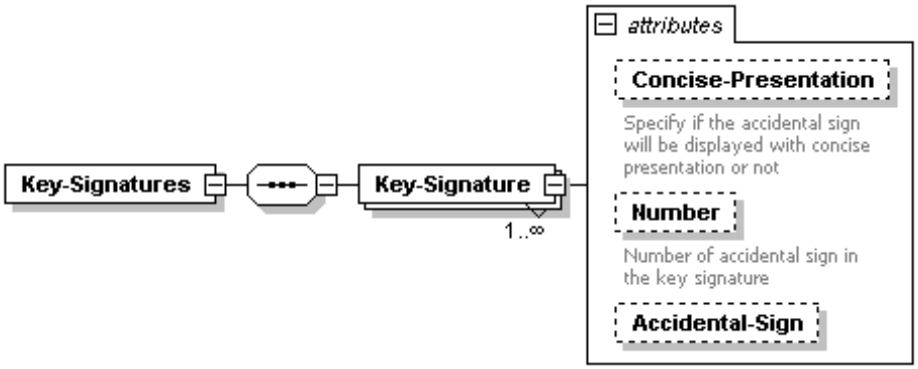

Figure 44 : Structure de l'élément Armure de Clé (Key-Signature).

Pour exprimer la signature rythmique des mesures, les partitions Braille utilisent deux cellules (supérieure et inférieure) précédées par le signe de nombre (indiquant que ceux qui vont suivre sont des nombres) au début de la combinaison.

La cellule supérieure est utilisée pour une signature rythmique à un seul chiffre tandis que les deux cellules sont utilisées pour une signature rythmique à deux chiffres.

La structure de l'élément signature rythmique est présentée dans la Figure 45.

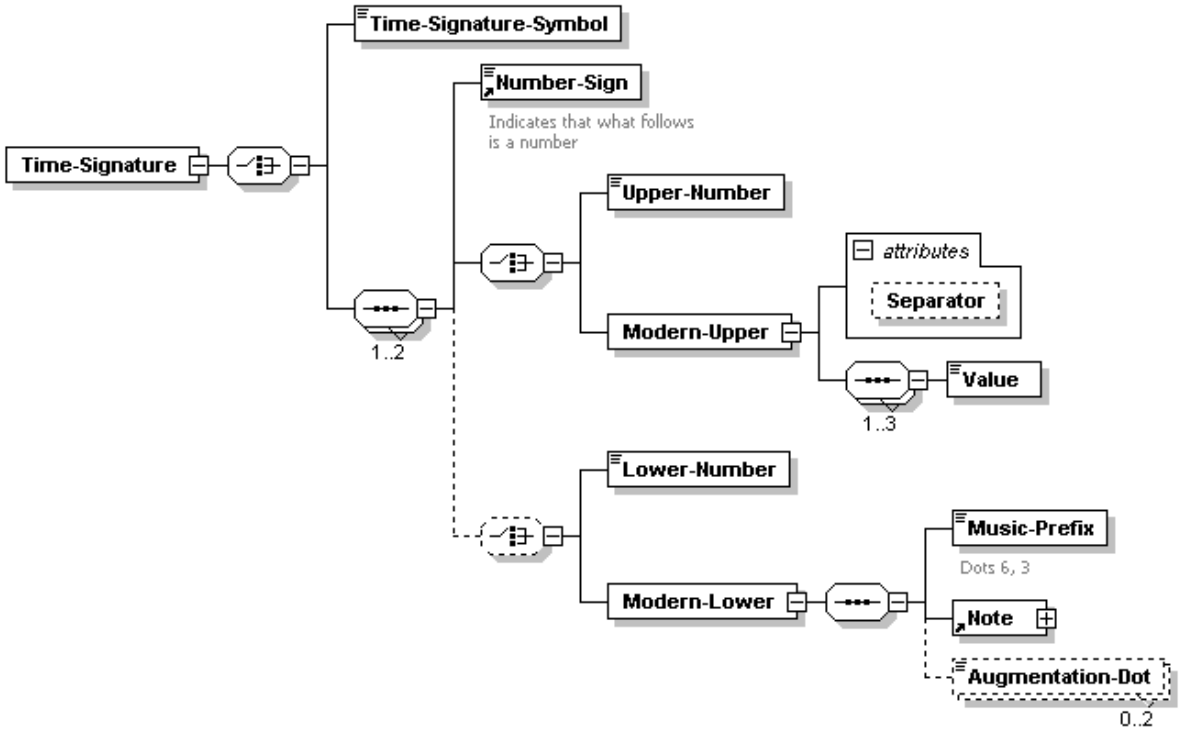

Figure 45 : Structure de la Signature Rythmique en BMML.

L'élément *Upper-Number* est l'équivalent de la valeur d'un temps (beats) en MusicXML tandis que *Lower-Number* est le nombre de temps dans une mesure (beat-type).

Pour afficher les informations telles qu'elles sont écrites dans les partitions en noir, l'élément *Time-Signature-Symbol* est utilisé pour enregistrer la valeur C (pour une mesure à 4 temps, valeur d'un temps = 4 et nombre de temps par mesure = 4) ou C barre (pour une mesure à deux temps, valeur d'un temps = 4 et nombre de temps par mesure = 2).

Un autre concept important est qu'en notation moderne, quand une note apparaît dans la signature rythmique (paragraphe 13-18 du nouveau manuel international de la musique Braille), elle est précédée par le préfixe de musique (point 6 suivi par 3) et la note C est utilisée pour représenter la valeur affichée en noir. L'élément Modern-Lower est utilisé pour mémoriser cette information.

De même, comme précisé dans le paragraphe 13-19 du nouveau manuel international de la musique Braille, la musique Braille dispose de deux signatures rythmiques modernes inhabituelles dont celle utilisant deux signatures de temps posées l'une après l'autre et celle utilisant plus d'une marque de valeur de temps. Dans le premier cas, les cellules supérieures et inférieures sont utilisées deux fois, une pour chaque signature rythmique, tandis que dans le second cas, l'élément Modern-Upper est utilisé.

Pour l'élément Clé (Key), l'attribut valeur (Value) prend sa valeur parmi les huit propriétés telles que G, G main gauche, F, F main droite, C, C en quatrième ligne, G avec signe d'octave au-dessus de l'armature de clé, ainsi que G avec signe d'octave au-dessus de l'armature de clé.

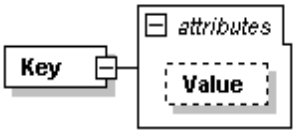

Figure 46 : Structure de l'élément Clé.

Dans la Figure 47, parmi les connecteurs de mesure, nous pouvons trouver des signes de répétition de mesure. Dans la musique Braille, il y a trois sortes de répétition de mesure : répétition de mesure, répétition de mesure éloignée et répétition de mesures numérotées. Chacune d'entre elles commence par le symbole de répétition (des points 2356).

Tout d'abord, dans le cas de répétition d'une mesure, l'élément *Number* indique le nombre de fois où la mesure actuelle doit être rejouée. Cet élément doit être précédé par le signe de nombre qui indique que le symbole suivant est un nombre. De même, pour indiquer une répétition de mesure éloignée, l'élément de *Backward-Mark* spécifie le point de début à partir duquel un certain nombre de mesures indiquées dans l'élément *Number* doivent être rejouées. Finalement, en termes de répétition de mesures numérotées, les numéros de mesures de début et de fin sont spécifiés après le symbole de trait d'union (*Hyphen-symbol*).

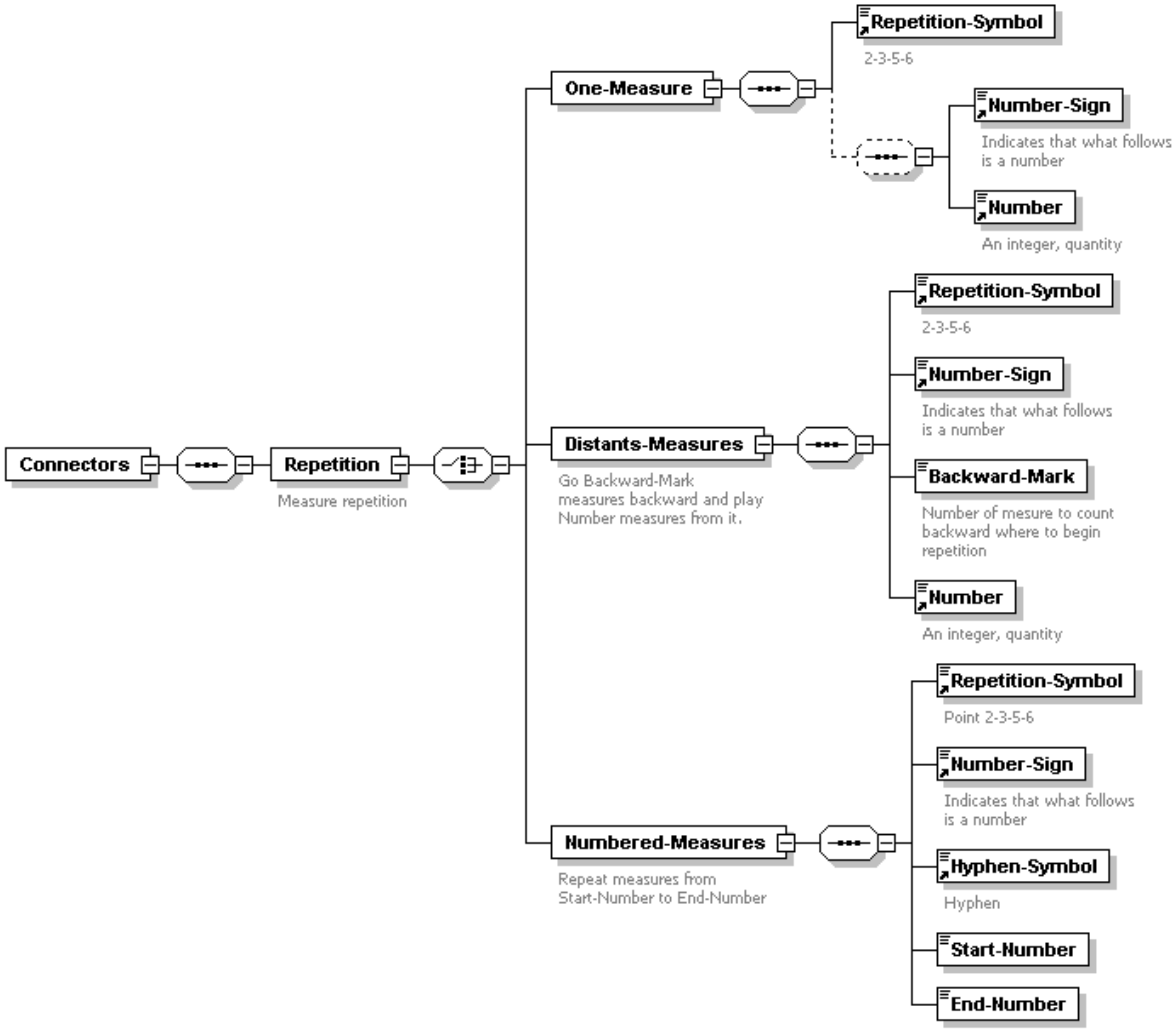

Figure 47 : Structure de l'élément connecteur de mesures.

Comme nous l’avons indiqué précédemment, dans une partition Braille, une mesure (*Measure*) est composée de séquences de *Notes* qui peuvent être connectées les unes aux autres par les connecteurs de notes. La Figure 48 montre qu’une mesure, identifiée par la propriété de *Number*, a une propriété *Type* qui, suivant sa valeur, spécifie si l'espace entre les mesures ou le retour à la ligne représentent la barre de mesure, si elle a une signe de barre de mesure Braille, si elle a une barre de mesure pointillée, une barre de mesure double à la fin de composition ou une barre de mesure double à la fin de section.

Dans la Figure 48, l’élément *Notes* est un ensemble de « *Note* » qui est précédé par des pré-descripteurs et suivi par des post-descripteurs. Ils correspondent à des éléments qui doivent être mis avant et après la note selon des règles Braille. De plus, les Notes peuvent être connectées l'une avec l'autre par des éléments de *Connector*.

En outre, dans une partition Braille, les notes et les silences sont représentés de la même façon. L'élément *Notes* est défini récursivement de sorte qu’une séquence de *Notes* puisse contenir un autre groupe de *Notes*. Ceci facilite l’expression des groupes rythmiques qui se chevauchent comme les triolets dans un triolet.

De surcroît, chaque note dans un accord est décrite par l'intervalle (*Interval*) qu'elle fait par rapport à une note de référence ou par l'élément *Accord*.

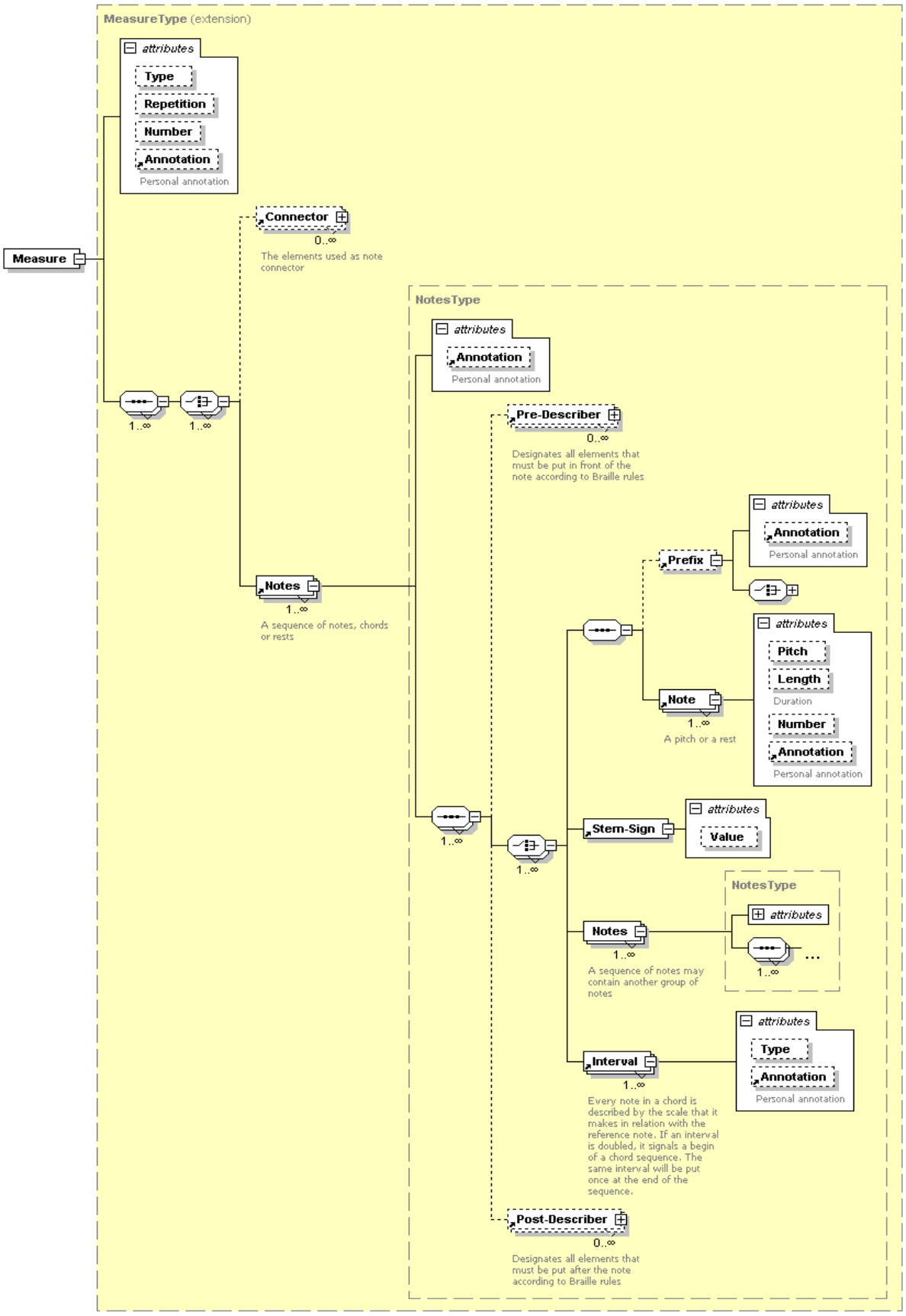

Figure 48 : Structure de l'élément Mesure.

Suivant les règles de la musique Braille, les sous-éléments de l'élément *Connector* sont utilisés pour signaler une relation ou une connexion entre deux notes. Parmi eux, nous pouvons citer : le symbole de la queue des notes, le symbole de copule, le symbole de trait d'union, le symbole de répétition et le symbole de dynamique.

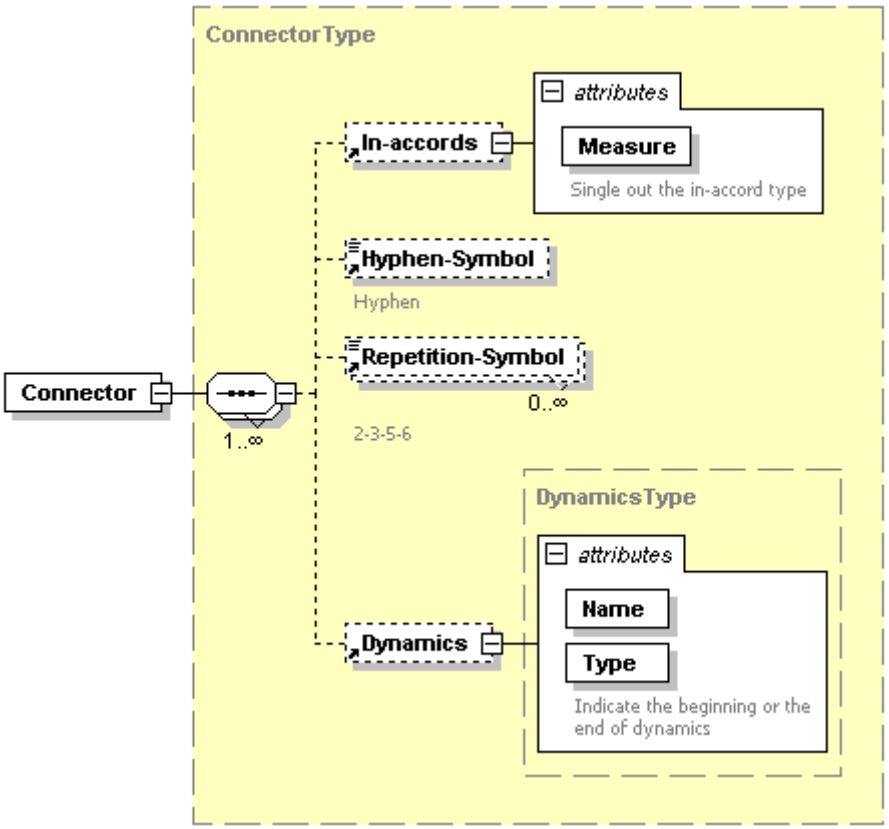

Figure 49 : Structure de l'élément Connector (Jessel et al, 2007).

La valeur d'un signe de queue de note peut être attribuée en choisissant une valeur parmi rond, rlanche, noire, croche, double croche et triple croche.

L'attribut *Measure* de l'élément copule (*In-accords*) spécifie s'il concerne une mesure entière ou une partie de mesure.

L'élément *Dynamics* a deux attributs qui sont *Name* et *Type* de la dynamique. L'attribut *Name* prend sa valeur parmi *crescendo, rinforzando, sforzando …* tandis que l'attribut de Type peut être *début* ou *fin*.

Il est important de signaler que l'utilisation de l'élément *Connector* facilitera l'analyse des partitions Braille par les logiciels de lecture de partition braille comme BMML Reader.

4.3.3.5. La partie Lyrique de BMML

Cette section traite la partie Lyrique de BMML schema.

Comme, une partie de solo vocale est transcrite comme une partie d'instrument à part la transcription des liaisons et l'insertion des textes vocaux (Nouveau Manuel International de

Musique Braille, 16-1), le lyrique est traité à deux niveaux. Tandis que le texte lyrique est géré à l'aide de l'élément *Lyrics*, *Post-describer* contient les liaisons qui indiquent qu'un texte est chanté sur deux ou plusieurs notes afin de garantir la synchronisation du texte avec la musique.

Dans BMML, le concept de lyrique est inspiré de SMR (Symbolic Music Representation) qui est présenté dans les Figure 50 et Figure 51.

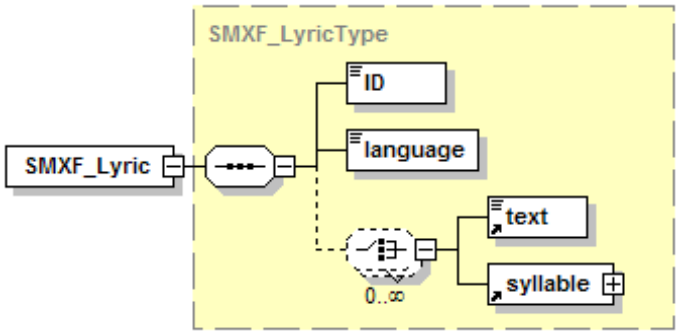

Figure 50 : Structure de l'élément Lyric dans SMR.

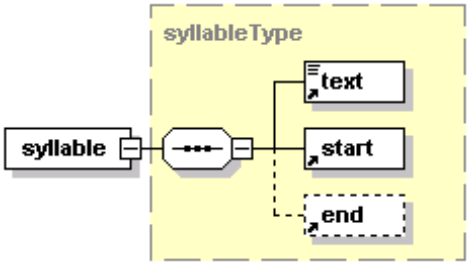

Figure 51 : Structure d'une Syllabe dans SMR.

Pour SMR, l'élément lyrique est un ensemble de textes d'une langue qui est subdivisé en plusieurs syllabes. Une syllabe peut être étendue sur plusieurs notes ou associée avec une seule note. Un élément *Syllable* peut contenir :

- un élément texte qui contient le texte de la syllabe,
- un élément *start* indiquant la note à partir de laquelle la syllabe s'applique,
- un élément optionnel *end* désignant la note à laquelle se termine la syllabe.

Cependant, l'élément lyrique, comme il a été défini par SMR n'est pas suffisant pour la représentation en Braille car, les cas suivants doivent être considérés :

- une note pour une syllabe,
- une note pour plusieurs syllabes,
- plusieurs notes pour une seule syllabe,
- une répétition de *syllables*, notion qui n'existe pas dans SMR.

Pour pouvoir prendre en compte ces éléments, nous avons choisi de représenter l'élément lyrique de BMML comme indiqué dans la Figure 52.

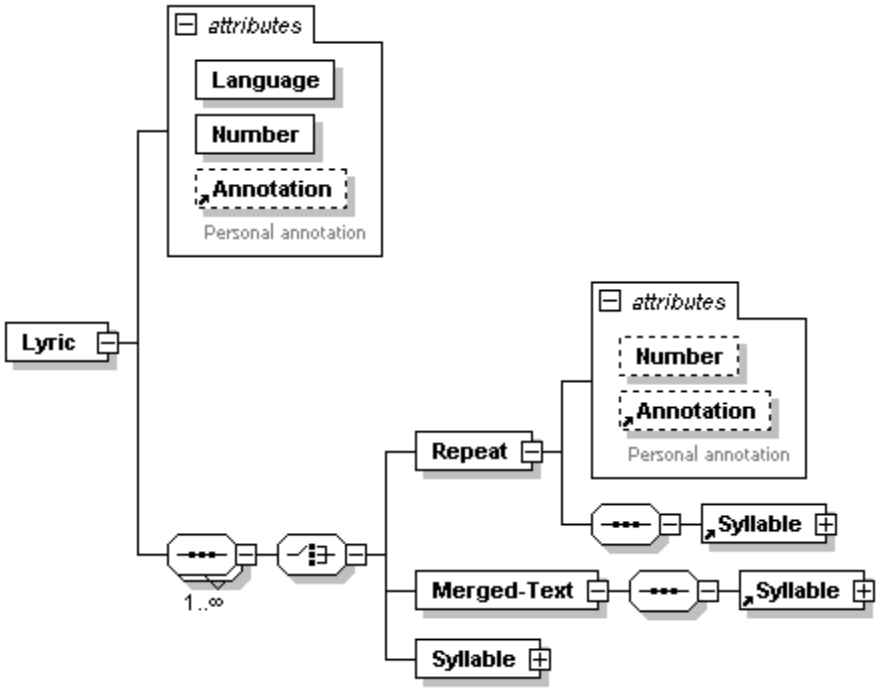

Figure 52 : Détails de l'élément Lyrics dans BMML

Les paroles peuvent être écrites dans plusieurs langues et chaque partie lyrique est référencée par son numéro. L'élément *Merged-text* (Texte fusionné) est utilisé pour exprimer deux ou plusieurs syllabes chantées sur une seule note.

L'élément *Repeat* (répétition) est utilisé pour enregistrer des mots ou des phrases qui sont répétés. L'attribut *Number* (nombre de répétitions) permet de définir le nombre de répétitions si celui-ci est supérieur à deux.

En ce qui concerne l'élément Syllabe, il a trois fils : *Text*, *Start* et *End*. L'élément *Text*, comme son nom l'indique, contient le texte de la syllabe tandis que *Start* et *End* se réfèrent à une note donnée dans la partition. La note est référencée non seulement par la section et mesure dans laquelle la note se trouve mais également par le numéro d'ordre de la note dans la mesure. La structure détaillée de l'élément syllabe est présentée dans la Figure 53.

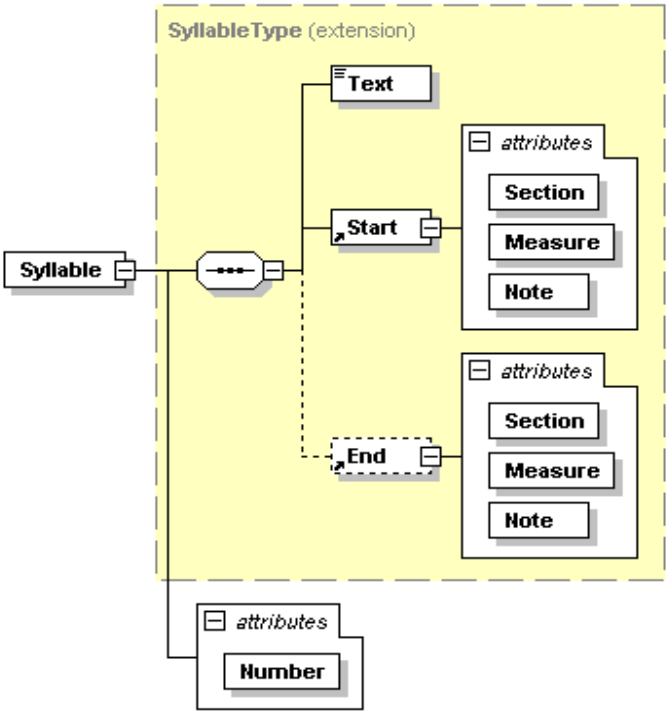

Figure 53 : Structure de l'élément Syllable dans BMML.

L'élément *Syllable* possède un attribut *Number*, par lequel il est référencé dans la partie Symbole d'accord.

4.3.3.6. Les accords

Dans les partitions Braille, les accords peuvent être décrits soit par une note suivie par des intervalles soit en utilisant les symboles d'accords de type Cmaj7, Dmin7 comme décrit dans la section suivante.

4.3.3.7. Accord décrit par symbole d'accord

L'élément *Chord-Symbols* est utilisé pour décrire des séquences d'accords en utilisant des symboles d'accords (comme CM, Dm7, G7) plutôt que des signes d'intervalles. Il accompagne la musique ou la partie lyrique ou les deux.

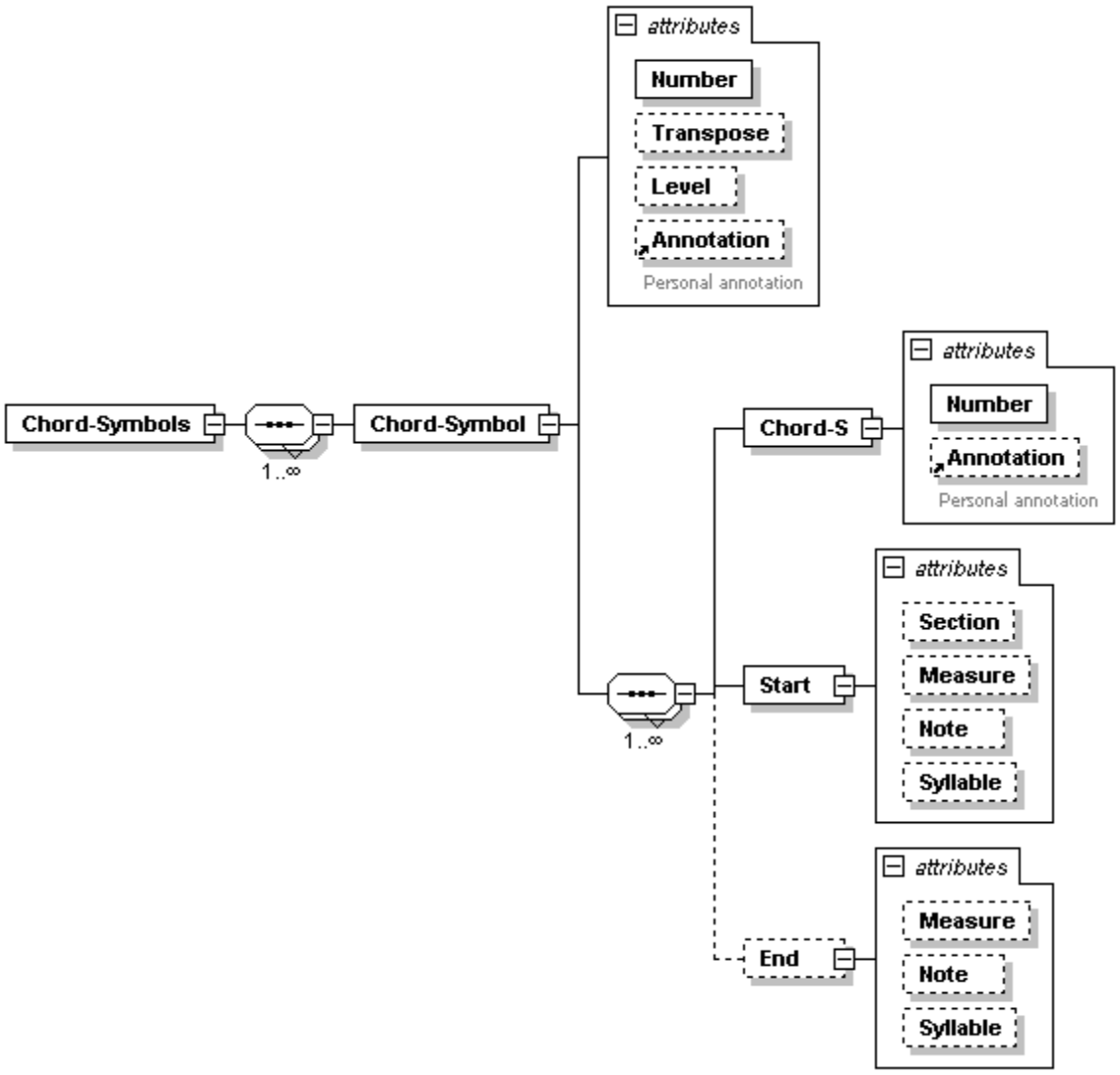

Figure 54 : Structure de l'élément Symbole d'accord (Chord-Symbol) dans BMML.

Différents types de séquences d'accords peuvent être écrits pour une chanson donnée. De ce fait, une chanson peut être écrite de plusieurs façons suivant le niveau musical voulu. Ainsi, des séquences d'accords écrites dans un niveau avancé peuvent substituer des accords basiques au niveau débutant. Dans ce cas, un attribut nommé *Level* (niveau) est ajouté à l'élément de Symbole d'accord (Chord-Symbol).

L'attribut *Transpose* sert à ajuster la tonalité de la chanson par rapport au symbole d'accord enregistré.

Chord-S est l'élément qui décrit le nom de l'accord. Par exemple Dm7, pour l'accord Ré mineur septième.

Les éléments *Start* et *End* contiennent respectivement le début et la fin de l'accord. Ils se réfèrent soit à une note donnée d'une mesure soit à une syllabe lyrique (dans le cas de symbole d'accord avec lyrique seulement).

4.3.3.8. Accord décrit par intervalle

Dans ce cas, chaque note dans un accord est décrite par l'intervalle qu'il fait avec la note de référence. Si un intervalle est doublé, cela signale un début de séquence d'accords. Le même intervalle (qui a été doublé) sera mis une seule fois à la fin de la séquence d'accords.

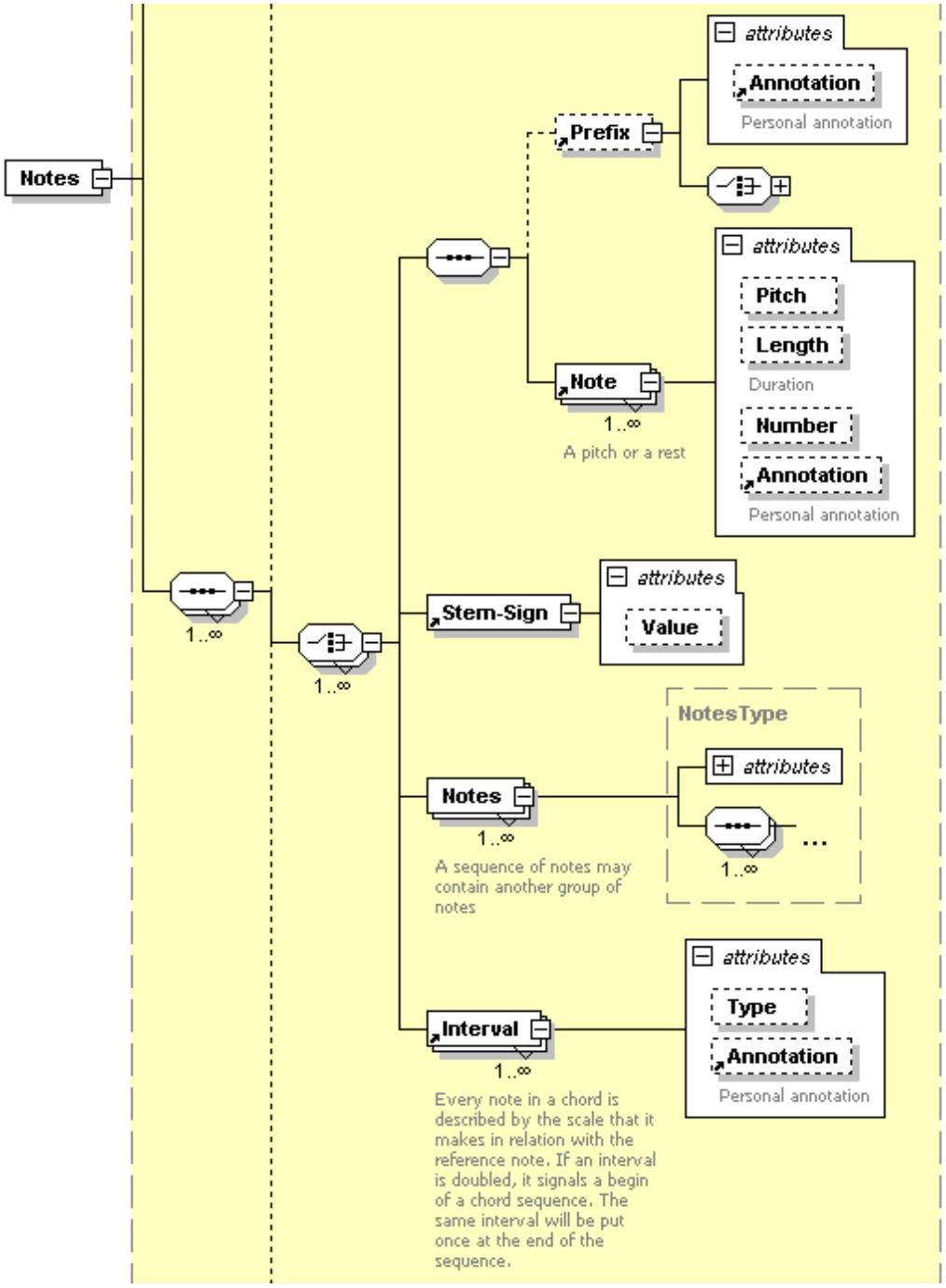

Figure 55 : Structure de l'élément Notes montrant l'élément Intervalle comme signe d'accords.

4.3.4. Illustration

La Figure 56 présente un exemple de séquence où un ensemble d'accords de tierce est répété.

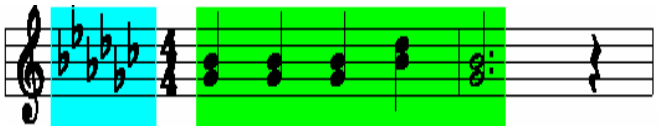

Figure 56 : Exemple de séquence d'accords répétés.

Voici le document XML correspondant à la partition, selon le schéma BMML.

```
<?xml version="1.0" encoding="UTF-8"?>
<Score xmlns="http://www.punctus.org/bsml"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.punctus.org/bsml/bmml.xsd">
        <Part Instrument="Piano" Number="1" Direction="Up" >
                <Section Number="1"  Movement="Moderato">
                        <Key Value="G-Clef"/>
                        <Key-Signatures>
                                <Key-Signature Number="6" Accidental-Sign="Flat" Concise-
                                Presentation="true"/>
                        </Key-Signatures>

                        <Measure Number="1">
                                <Notes>
                                    <Pre-Describer>
                                        <Octave  Number="4" Display="true"/>
                                    </Pré-Describer>
                                    <Note Pitch="G" Length="Quarters-64ths"/>
                                    <Interval Type="Third"/>
                                    <Interval Type="Third"/>
                                    <Note Pitch="G" Length="Quarters-64ths"/>
                                    <Note Pitch="G" Length="Quarters-64ths"/>
                                    <Note Pitch="B" Length="Quarters-64ths"/>
                                  </Notes>
                        </Measure>

                        <Measure Number="2">
                                <Notes>
                                    <Note Pitch="G" Length="Halves-32nds" />
                                        <Post-Describer>
                                                <Augmentation-Dot/>
                                         </Post-Describer>
                                         <Interval Type="Third"/>
                                    <Note Pitch="Rest" Length="Quarters-64ths"/>
                                </Notes>
                        </Measure>
                </Section>
        </Part>
</Score>
```

La correspondance Braille est donnée par une table permettant de reconstituer les caractères Braille. Pour l'indication d'armure (nombre de bémols à la clé), le nombre aura une correspondance en Braille ainsi que le bémol. L'attribut *concise-presentation = vrai* indique que la présentation respecte la règle Braille d'abréviation qui est de dire que dans le cas où il y a plus de 4 altérations leur nombre est indiqué par une information numérique.

Chaque note (Hauteur + Durée) a une correspondance Braille unique. A l'intervalle de tierce correspond le caractère Braille contenant les points 3, 4 et 6.

Cette table de correspondance est utile pour conserver des indications non conventionnelles utilisées par certains centres de transcription.

4.3.5. Annotation sémantique d'une partition BMML

L'annotation sémantique des partitions musicales Braille se fait en remplissant les métadonnées musicales pré-intégrées dans la structure des partitions Braille avec les termes des concepts musicaux issus d'une ontologie musicale (Raimond et al., 2007).

En outre, des commentaires peuvent être insérés par l'auteur ou toute personne qui souhaite apporter des précisions sur un passage dans la partition.

4.3.6. Recherche et accès à des partitions BMML

Les besoins d'accès à l'information musicale pour un non voyant dépendent de son statut et surtout pourquoi il veut avoir accès à cette information. Un élève ou un professeur de musique, un musicien professionnel ou un amateur sont définis comme des profils ayant des besoins différents mais il est possible d'identifier d'autres critères permettant de définir la présentation la mieux adaptée. Le niveau musical, le niveau de connaissance du Braille ainsi que le niveau de handicap sont à prendre en compte. Les outils logiciels et matériels utilisés et développés dépendent de la tâche à effectuer, du type de profil mais également du niveau de l'utilisateur.

Pour qu'une partition musicale soit accessible aux utilisateurs, il faut donc pouvoir agir sur ces différents critères et trouver dans la représentation de la partition tous les éléments permettant de les prendre en compte. Un premier travail a été réalisé au niveau des codes musicaux existants pour coder la musique et analyser du code musical Braille.

Prenons une série d'exemples spécifiques de tâches de consultation d'une partition musicale ainsi que les principes à mettre à œuvre pour répondre à ces besoins. L'expression des besoins concernant l'accès à une partition donnée est décrit par une série de scénarii, l'hypothèse de travail ici est que la consultation s'effectue sur une partition préalablement trouvée et codée à l'aide d'un langage spécifique.

Un élève de musique d'une école de musique, niveau élémentaire en solfège, ayant une connaissance très élémentaire du Braille pourra demander une lecture du nom des notes et de leur durée. Un pianiste débutant pourra vouloir des partitions avec ou sans les liaisons, nuances ... pour qu'il puisse plus facilement les mémoriser étape par étape. Un autre étudiant pourra vouloir analyser la même partition sans les doigtés.

Avant que ces besoins de lecture des partitions ne soient pris en compte, il est nécessaire que celles-ci soit numérisées et éventuellement produites en Braille.

Dans le projet Contrapunctus**,** nous avons contribué à la spécification, la conception et au développement du langage de description des partitions musicales Braille BMML en vue non

seulement de pouvoir enregistrer et interpréter des partitions Braille, mais également de pouvoir les publier en ligne et les rechercher.

Ainsi, nous avons conçu dans BMML des métadonnées spécifiques à tous niveaux structurels et sémantiques des documents Braille (partie, section, mesure,…). Ces métadonnées ont été élaborées afin de permettre d'effectuer des recherches ultérieures sur ces documents. A partir de ces métadonnées, divers types de recherches peuvent être effectués, à savoir, les recherches par titre, auteur, compositeur, date de création, date de production, genre musicale…

Du point de vue purement musical, la partie dédiée à la description des accords d'une partition permettra d'effectuer des analyses harmoniques de la partition. Ceci permettra par exemple aux musiciens de rechercher les partitions contenant des cadences musicales (progressions ou suites d'accords) données.

De plus, tous les modes de recherches utilisés sur des documents textes peuvent être appliqués aux documents musicaux Braille puisque BMML comprend aussi la partie *Lyrics* dédiée à la parole associée à la partition en question.

Chapitre 5. CONCLUSION ET PERSPECTIVES

La conception d'un système de recherche d'information efficace et satisfaisant pour l'utilisateur ne se limite pas à l'amélioration des composantes d'un moteur de recherche. La structure des documents de la collection ainsi que des connaissances qui y sont véhiculées doivent également être prises en compte.

En effet, afin de pouvoir distinguer, parmi les documents de la collection, ceux qui sont potentiellement pertinents par rapport à la requête de l'utilisateur, il est important de comprendre d'une part les informations contenues dans les documents du point de vue structure et sémantique, et d'autre part celles de la requête de l'utilisateur. Cependant, dans les SRI actuels, la sémantique des contenus n'est pas prise en compte car les index qui sont utilisés pour la description des documents sont plutôt basés sur la statistique d'apparition des occurrences des termes dans les documents.

Dans le cadre de cette thèse, nous avons répertorié et analysé différents paramètres qui peuvent influencer les résultats d'une recherche de documents dans un SRI. Parmi ces paramètres, la structuration des documents, les connaissances véhiculées dans les documents, leurs contextes d'utilisation ainsi que les informations personnelles relatives à l'utilisateur ont été étudiés. Un système d'information qui a pour objectif de fournir aux utilisateurs les documents répondant à leurs requêtes doit tenir compte de ces différents paramètres pour atteindre cet objectif.

Dans la littérature, des modèles conceptuels et opérationnels des documents ont été développés afin de pouvoir comprendre et analyser les contenus des documents. Néanmoins, il manque d'une part la considération de l'utilisateur et du contexte dans lequel il veut utiliser le document et d'autre part la description sémantique du contenu des documents afin de mieux comprendre les messages véhiculés dans ceux-ci.

Ainsi, afin de prendre en compte d'un côté la structure et la sémantique des documents et d'un autre côté la sémantique des requêtes des utilisateurs et le système d'appariement requête-document, un méta-modèle de représentation des documents a été proposé en faisant abstraction des particularités et spécificités des divers domaines d'applications. Ce modèle générique sert de représentation multi-facette de base applicable à divers domaines et est basé sur l'utilisation d'ontologie pour l'aspect sémantique. L'instanciation de ce méta-modèle pour une application particulière permet d'obtenir un modèle adapté et conforme à celui-ci. Nous avons proposé trois instanciations de notre méta-modèle dans des applications différentes dont : l'apprentissage en ligne, la musique Braille et la maintenance automobile.

Dans le cadre de l'apprentissage en ligne, malgré l'existence des normes comme SCORM, LOM, IMS-LD qui visent respectivement à la structuration des objets pédagogiques, à la description de leurs contenus et à leur agencement, il manque la prise en compte de la notion d'approche pédagogique pour l'apprentissage. Ce problème est constaté aussi bien au niveau de la présentation des ressources pédagogiques qu'au déroulement de l'apprentissage. De plus, le suivi des connaissances acquises par les apprenants n'a pas été pris en compte par les plateformes d'apprentissage en ligne de la littérature.

Pour remédier à ces manques, en partant du modèle générique de représentation des documents, un modèle de représentation multi-facette des objets pédagogiques a été instancié (Hernandez et al, 2008). Ce modèle contient, outre les structurations des objets pédagogiques et les sémantiques de leurs contenus, les contextes qui englobent les différents scénarii et tâches d'utilisation des objets pédagogiques. Les différents rôles pédagogiques comme celui de l'enseignant et celui de l'apprenant sont aussi pris en compte pour élaborer des scénarii d'apprentissage utilisant les objets pédagogiques. Cette représentation permet de faciliter la recherche sémantique des objets pédagogiques en vue de les réutiliser dans d'autres objets pédagogiques ou activités d'apprentissage pour la conception de nouveaux cours ou de scénarii pédagogiques. Aussi, les différents contextes d'utilisation des objets pédagogiques sont pris en compte pendant la phase de recherche car ces objets sont indexés non seulement par les termes de l'ontologie de thèmes du domaine de l'étude mais également celle des tâches décrivant leurs utilisations. La prise en compte du contexte d'utilisation des documents et leur indexation sémantique permettent également de suivre et de contrôler les différentes connaissances acquises par les apprenants au cours de leurs activités d'apprentissage. Dans notre approche, à tout moment de leurs apprentissages, les profils des apprenants sont dressés sous forme d'ontologies personnelles. Chaque apprenant peut donc être évalué en comparant son ontologie personnelle avec celle du domaine de l'étude d'un cours dans une formation donnée. Ceci permet au système de personnaliser les scénarii d'apprentissage suivant les pré-requis de chaque objet pédagogique, les phases d'apprentissage auxquelles il est associé et le profil de chaque utilisateur. Pour valider notre travail, une plateforme d'apprentissage en ligne multimédia (PALM) a été développée en tenant compte de toutes nos propositions dans ce domaine d'application.

Nous avons également analysé un cas particulier de documents que sont les partitions musicales (dans le cadre d'un projet européen). En effet, les besoins en recherche et d'échange de partition musicale sur la toile ne cessent de s'accroître. Des normes et standards (SMDL, SMR) ont été élaborés afin d'encoder les partitions musicales d'une part et de permettre les échanges de données musicales d'autre part. Des formats d'encodage (MusicXML, NIFFML, MIDI) issus de ces standards ont été réalisés pour l'encodage des partitions musicales, cependant aucun de ces formats n'a pris en compte la partition musicale Braille. Ainsi, les malvoyants n'ont pas accès aux ressources musicales électroniques. Ceci nous amène à instancier notre modèle de représentation multi-facette des documents dans le domaine de la musique Braille et de proposer en même temps un format d'encodage des partitions musicales Braille qui tienne compte non seulement des normes sur les documents musicaux mais également des particularités de la musique Braille et ses contextes d'utilisation. Ce modèle comprend la présentation dynamique et statique des partitions musicales Braille, leur description thématique par une ontologie musicale, la description de leurs contenus par des métadonnées ainsi que la description de leurs contextes d'usages.

Cette prise en compte de la structuration des documents musicaux Braille et du contexte de l'utilisateur a permis de personnaliser la présentation du document suivant le profil de l'utilisateur. Ainsi, un même document Braille peut être présenté différemment à l'utilisateur selon qu'il soit apprenant ou enseignant, voyant, mal voyant ou non voyant, musicien débutant ou avancé. Ainsi, ce modèle de représentation de documents facilite aux non voyants et malvoyants non seulement la recherche et l'accès aux ressources musicales Braille mais également l'apprentissage de la musique Braille.

Enfin, l'une des cadres applicatifs pour lequel nous avons instancié notre méta-modèle est la mise en place d'un système de recherche de documents appelé Dynamo (pour Dynamic Ontology for Information Retrieval). Ce projet, soutenu par l'ANR, nous a permis de mettre en place et d'évaluer les études que nous avons menées durant cette thèse.

Dynamo est un système qui permet d'indexer sémantiquement les documents d'un domaine à l'aide d'une ontologie de domaine, sachant que les documents, comme les ontologies qui représentent les connaissances de ces domaines, peuvent évoluer dans le temps. L'architecture du système ainsi que le modèle de données que nous avons conçus ont été élaborés en prenant en compte cet aspect dynamique au niveau des documents et des ontologies de domaine. Le modèle de représentation multi-facette que nous avons proposé permet cette prise en compte tout en étant flexible et adaptable à des contextes et applications variées.

Dans Dynamo, les documents ont été représentés par une facette qui décrit la sémantique des contenus des documents à l'aide d'une ontologie de domaine, une facette de description de la structure des documents et une facette qui décrit les usages des documents. Le système permet d'indexer et de rechercher tout domaine d'application. Dans le cadre de cette thèse, la collection utilisée, qui est celle du partenaire industriel ACTIA, est relative au domaine de la maintenance automobile. Ce système annote les documents (respectivement les requêtes) de l'utilisateur à l'aide des graphes de concepts ou des concepts isolés qui expriment les connaissances thématiques véhiculées aussi bien dans les documents (respectivement dans les requêtes), à travers les ontologies qui représentent les connaissances du domaine et les index qui représentent les connaissances véhiculées dans chaque document. Ainsi, le processus d'appariement entre la requête et les documents consiste à comparer l'ensemble des concepts qui les annotent à l'aide des fonctions de similarité sémantique. Ces fonctions de similarité sémantique sont basées sur les mesures de similarité conceptuelle qui expriment la corrélation ou la similitude entre concepts. Diverses mesures de similarité conceptuelle ont été développées (Wu et Palmer, 1994), (Resnik, 1995), (Jiang et Conrath, 1997) (Lin, 1998) (Leacock et Chodrow, 1998). Néanmoins, il manque une mesure de similarité conceptuelle basée sur la sémantique des concepts et adaptée à divers types d'application. Ainsi, nous avons proposé des fonctions de similarité conceptuelle qui tiennent compte de la hiérarchie des concepts qui reflètent les proximités sémantiques entre concepts. A partir de ces mesures de similarité conceptuelle que nous avons proposées, nous avons développé une fonction de similarité sémantique générique qui évalue la similarité entre les documents et la requête par une combinaison des similarités de concepts de même type.

Les évaluations que nous avons réalisées dans le cadre du projet Dynamo, moyennant les ontologies et les collections des entreprises partenaires du projet, ont montré que les mesures de similarité conceptuelle baptisée ProxiGénéa (pour proximité généalogique) que nous avons proposées sont plus efficaces que celle de Wu et Palmer en termes de précision moyenne globale d'une part et du rapport rappel et précision d'autre part.

Indépendamment de ces cadres d'applications, l'une des problématiques des SRI, qui sont en général destinés à être utilisés sur la toile ou dans des environnements disposant des collections de documents hautement dynamiques, est de garder la cohérence entre les documents et les index. Divers travaux ont été élaborés afin de prendre en compte les dynamiques des documents et de gérer leurs impacts sur les index. Des problématiques comme la disponibilité des documents et les délais de mise à jour des index restent à résoudre. Pour solutionner ce problème, nous avons conçu un modèle d'indexation dynamique à base d'ontologies ainsi que les algorithmes de mise à jour d'index correspondants (Hubert et al., 2009). Ce travail a permis la disponibilité des documents ainsi que les index décrivant ces documents.

En conclusion, comme le besoin en information diffère d'un utilisateur à l'autre et d'un domaine à l'autre, pour qu'un SRI soit efficace et réponde avec satisfaction aux besoins des utilisateurs, des informations importantes autour des documents et de l'utilisateur doivent être prises en compte en même temps. Parmi ces informations nous pouvons citer : les structures

des documents, la sémantique de leurs contenus, leurs contextes d'utilisations ainsi que les profils des utilisateurs.

Plusieurs perspectives sont envisagées pour poursuivre ce travail. Concernant les mesures de similarité conceptuelle, nous envisageons l'amélioration des formules que nous avons proposées en considérant les axiomes exprimés sous forme de description logique dans les ontologies. Ceci permettra d'une part d'être plus précis en matière de RI sémantique, et d'autre part d'étendre notre champ de recherche dans les stratégies d'alignement d'ontologies en vue de la gestion des connaissances hétérogènes.

En ce qui concerne le passage à l'échelle, comme cela est nécessaire lorsque l'on s'intéresse au web, nous envisageons l'intégration dans notre système les techniques de compression de données (Witten et al., 1994) tout en essayant de rester efficace en termes de précision et de temps de réponse. De plus, nous projetons une classification et une analyse multidimensionnelle des documents au moment de leur insertion dans la collection (Khrouf et Soulé-Dupuy, 2005).

Contrairement à l'approche sac de mots, le vocabulaire utilisé lors de l'indexation peut être amené à varier indépendamment des documents. En effet, une ontologie en tant que représentation de connaissances d'un domaine évolue pour représenter au mieux le domaine. L'ontologie qui a servi de base de référence pour le choix des termes d'indexation des documents peut donc évoluer. La cohérence de l'ontologie et de l'indexation des documents n'est alors plus assurée. Dans ce cas, il est donc important de considérer la mise à jour de l'indexation consécutive à une modification du vocabulaire de référence, cela afin de maintenir une cohérence entre les documents et le vocabulaire d'indexation en vue de faciliter la recherche des documents. L'une des perspectives à nos travaux de recherche consiste à prendre en compte cette évolution et ses impacts sur l'indexation des documents.

En outre, pour garder la cohérence entre ontologie de domaine et documents de la collection, nous envisageons l'intégration dans nos recherches des travaux réalisés d'une part dans (Bourigault et al., 2004) sur la construction automatique d'ontologie à partir des textes et d'autre part dans (Chrisment et al., 2006) sur la mise à jour d'une ontologie de domaine à partir de l'analyse de nouveaux documents du domaine.

REFERENCES

Abel, M.,H., Lenne, D., Moulin, C. et Benayache A., « Gestion des ressources pédagogiques d'une e-formation ». *Document Numérique* 7(1-2), p. 111-128, (2003).

Andreasen T., Bulskov H., Knappe R., « Similarity for Conceptual Querying », *Proceedings for the 18th International Symposium on Computer and Information Sciences,* pp 268-275, (2003).

Aufaure M. A., Soussi R., and Baazaoui H., « SIRO: On-line semantic information retrieval using ontologies ». *2nd International Conference on Digital Information Management*, ICDIM'07, p. 321-326, (2007).Aussenac-Gilles, N., Mothe, J., « Ontologies as Background Knowledge to Explore DocumentCollections», *Recherche d'Information Assistée par Ordinateur (RIAO)*, p. 129-142, (2004).

Baeza-Yates, R.A., Navarro, G.: « Block addressing indices for approximate text retrieval». *Journal of the American Society on Information Systems* 51(1), p. 69–82, (2000).

Baranyi, P., Gedeon, T.,D., Koczy, L.T., « Intelligent information retrieval using fuzzy approach »., IEEE International Conference on Systems, Man, and Cybernetics Volume 2, p.1984 – 1989, vol.2. (1998).

Baziz, M., Boughanem, M., Aussenac-Gilles, N., Chrisment, C., « Semantic Cores for Representing Documents in IR », *In Proceedings of the 20th ACM Symposium on Applied Computing*, p. 1020-1026, ACM Press ISBN: 1-58113-964-0, (2005).

Benjamins, R., Fensel D., Decker D., Gomez Perez A., (KA)2, « Building ontologies for the internet : a mid-term report », *Proceedings of the International Workshop on Ontological Engineering on the Global Information Infrastructure*, pp 1-24, (1999).

Berry, M., W., Dumais, S., T., O'Brien, G., W.. «Using linear algebra for intelligent information retrieval». *SIAM Rev. 37, 4, p.* 573-595, (1995).

Berners-Lee, T., Hendler J., Lassila O., « The Semantic Web », *Scientific American*, p. 28-37, (2001).

Bourda, Y., « Objets pédagogiques, vous avez dit objets pédagogiques ? », *Actes du congrès, Gutenberg*, p.39-40, (2001).

Bourigault, D., Aussenac-Gilles, N., Charlet, J., «Construction de ressources terminologiques ou ontologiques à partir de textes : un cadre unificateur pour trois études de cas». *Revue d'Intelligence Artificielle (RIA), Numéro spécial sur les techniques informatiques de structuration de terminologies*, M. Slodzian (Ed.), Hermès, Paris, Vol. 18, N. 1/2004, p. 87-110, (2004).

Boutemedjet, S. Web Sémantique et e-Learning. *Cours FT6261*. (2004).

Bouzeghoub, A., Defude B., Duitama J.F., Lecocq C. « Un modèle de description sémantique de ressources pédagogiques basé sur une ontologie de domaine ». *Sticef*, vol. 12, (2005).

Büttcher, S. et Clarke, C., 2006. « A Hybrid Approach to Index Maintenance in Dynamic Text Retrieval Systems ». *Book Advances in Information Retrieval, Category* : *Performance and Peer-to-Peer Networks*, Springer Berlin / Heidelberg, Volume 3936. p. 229-240, (2006).

Castells, P., Fernandez M., Vallet D., « An Adaptation of the Vector-Space Model for Ontology-Based Information Retrieval ». *IEEE Transactions on Knowledge & Data Engineering*, Vol. 19 Issue 2, p. 261-272, (2007).

Chang, B., Ham D.H., Moon D. S., Choi Y. S., Cha J., «Using Ontologies to Search Learning Resources». *Book Series Lecture Notes in Computer Science*, Éditeur Springer

Berlin / Heidelberg ISSN 0302-9743, *Book Computational Science and Its Applications* – ICCSA,. Subject Collection Computer Science, p. 1146-1159, (2007).

Cho, J., Garcia M. H.: « The evolution of the web and implications for an incremental crawler ». In: *26th Intl. Conf. on Very Large Data Bases*, p. 200-209, (2000).

Chrisment, C., Hernandez, N., Hubert, G., Mothe. J., « Mise à jour d'une ontologie de domaine à partir de l'analyse de nouveaux documents du domaine pour l'indexation de documents »., *Information - Interaction - Intelligence*, Cépaduès Editions, Numéro spécial *Textes et ressources terminologiques et/ou ontologiques: évolution et maintenance*, Vol. Hors-série, p. 53-83, (2006).

Cutting D. R., Pedersen and J. O.. «Optimization for Dynamic Inverted Index Maintenance». *13th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, USA ACM Press. p. 405-411, (1990).

Collins, A., Loftus, E., «A Spreading Activation Theory of Semantic Processing». *Psychological Review*, vol. 82, p. 407-428, (1975).

Croft, W., B. Harding, S., M., Callan, J., P., « The INQUERY Retrieval System ». DEXA, p. 78-83, (1992).

De La Passardière, B, Jarraud P. « ManUeL, un profil d'application du LOM pour C@mpuSciences ». *Sticef*, vol. 11, p. 11-57. (2004).

Deerwester, S., C., Dumais, S., T., Landauer, T., K., Furnas, G., W., Harshman, R., A., «Indexing by Latent Semantic Analysis». *Journal of the American Society of Information Science*, Vol. 41:6, 391-407, (1990).

Dekang, L., « An information-theoretic definition of similarity ». *Proceedings of the 15th International Conference on Machine Learning*, p. 296–304. Morgan Kaufmann. (1998).

Desmontils, E. Jaquin, C., « Indexing a Web site with a terminology oriented ontology », *The Emerging Semantic Web*, I. Cruz S. Decker, J. Euzenat, D.L. McGuinness (Eds.), IOS Press, ISBN 1-58603-255-0, p. 181-197, (2002).

Dumais, S., « Latent Semantic Indexing (LSI) ». *TREC-3*, (1994).

Duval E., Sutton S. et Weibel, S.,L., « Metadata Principles and Practicalities », *D-Lib Magazine* 8(4), (2002).

Encelle, B., Jessel, N., Mothe, J., Ralalason, B., J., V., BMML: « Braille Music Markup Language », *International Conference on Internet Computing (ICOMP)*, Las Vegas, USA, (2008).

Euzenat J., « Eight questions about semantic Web annotations », *IEEE Intelligent systems* 17(2), p. 55-62, (2002).

Fage C. «Vous avez dit SCORM». *eLearning Agency, p.* 1. /. 14, (2005).

Foltz, P., W., «Using Latent Semantic Indexing for information filtering». CACM, p. 40-47, (1990).

Furnas, G., W., Landauer, T.,K., Gomez, L., M., Dumais, S.T., «The Vocabulary Problem in Human-System Communication», *Communications of the ACM 30*, p. 964-971, (1987).

Galambos L., « Dynamic Inverted Index Maintenance ». *World Academy Of Science, Engineering and Technology*, vol. 11, ISSN 1307-6884, p. 171-176, (2006).

Gasevic D. et Hatala M. « Searching context relevant learning resource using ontology mappings ». *International Workshop on Applications of Semantic Web Technologies for E-learning* (SW-EL), Winston-Salem State University, (2005).

Gligorov R., ven Kate W., Aleksovski Z., van Harmelen F., « Using Google Distance to Weight Approximate Ontology Matches ». *Proceedings of the 16th international conference on World WideWeb*, pp 767 - 776, 2007.

Guarino N, Masolo C., Vetere G., « OntoSeek: Content-Based Access to the Web », *IEEE Intelligent Systems*, 14 (3), pp 70-80, (1999).

Guha, R., V., McCool, R., Miller, E., « Semantic search », *In Proceedings of the 12th International World Wide Web Conference*, p. 700-709, (2003).

Haarslev, V. et Möller, R., Août 2001. « Description of the racer system and its applications ». *International Workshop on Description Logics*. Stanford, Californie, p. 132-141, (2001).

Haav H. M., Lubi T.L., « A Survey of Concept-based Information Retrieval Tools on the Web », *Proceedings of the 5th East-European Conference ADBIS*, Vol 2, p. 29-41, (2001).

Harman D., « Relevance Feedback Revisited », *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 1-10, (1992).

Hernandez N. « Ontologies de domaine pour la modélisation du contexte en Recherche d'Information ». *Thèse de l'Université Paul Sabatier* (2005).

Hernandez, N., Mothe, J., Ralalason, B., Stolf, P., «Modèle de représentation sémantique des documents électroniques pour leur réutilisabilité dans l'apprentissage en ligne». *Colloque International sur le Document Électronique*, Fribourg (Suisse), EUROPIA, p. 181-198, (2006).

Hernandez, N., Mothe, J., Chrisment, C., Egret, D., «Modeling context through domain ontologies». *Journal of Information Retrieval, Springer, Numéro spécial Contextual Information Retrieval Systems*, v*ol. 10 N. 2*, p. 143-172, (2007).

Hernandez, N., Mothe, J., Ralalason, B., Ramamonjisoa, B., Stolf, P.: « A Model to Represent the Facets of Learning Objects ». *Interdisciplinary Journal of E-Learning and Learning Objects, Informing Science Institute, Santa Rosa - USA, vol. 4*, p. 65-82, (2008).

Horrocks, I., « The FaCT system », *Automated Reasoning with Analytic Tableaux and Related Methods : International Conference Tableaux'98, number 1397 in Lecture Notes in Articial Intelligence*. Springer-Verlag, p. 307-312, (1998).

Hubert, G., Mothe, J., « An adaptable search engine for multimodal information retrieval », *Journal of American Society for Information Science and Technology*, Wiley, Vol. 60 N. 8, p. 1625-1634, (2009).

Hubert, G., Mothe, J., Ralalason B., Ramamonjisoa, B., « Modèle d'indexation dynamique à base d'ontologies », *Conférence francophone en Recherche d'Information et Applications*, Belambra de la Presqu'île de Giens, dans le Var, Laboratoire des Sciences de l'Information et des Systèmes (LSIS), p. 169-184, (2009).

Ingwersen, P., Järvelin, K., « The Turn: Integration of Information Seeking and Retrieval in Context » , *The Information Retrieval Series*, Springer-Verlag New York, Inc., Secaucus, NJ, (2005).

Jiang, J.,J., Conrath, D., W., « Semantic Similarity based on Corpus Statistics and Lexical Taxonomy ». *Proceedings of the International Conference on Research in Computational Linguistics (ROCLING)*. Taiwan, (1997).

Karvounarakis, G., Alexaki, S., Christophides, V., Plexousakis, D., Scholl, M., « Rql: a declarative query language for rdf », *In Proceedings of the 11th International World Wide Web Conference*, p. 592-603, (2002).

Khrouf, K., Soulé-Dupuy, C.,. « DocWare : Vers l'entreposage et l'analyse multidimensionnelle de documents », *Conférence en Recherche d'Information et Applications (CORIA'05), Grenoble - France*, IMAG Ed., ISBN : 2-9523810-0-3, p. 405-420, (2005).

Kim H., Park C. S., Park J. Y., Jung B., Lee Y. J., « A Multimedia Content Management and Retrieval System Based on Metadata and Ontologies ». *IEEE International Conference on Multimedia and Expo*, p. 556 – 559, (2007).

Kiryakov, A., Popov, B., Terziev, I., Manov, D., Ognyanoff, D.« Semantic annotation, indexing, and retrieval », *Journal of Web Semantics*, 2(1), (2004).

Köhler J., Philippi S., Specht M., Rüegg A., « Ontology based text indexing and querying for the semantic web ». *Knowledge-Based Systems*, Vol 19, Issue 8, p. 744 – 754, (2006).

Knight C., Gasevic D., Richards, G. « Ontologies to integrate learning design and learning content». *Journal of Interactive Media in Education (07)*, (2005).

Koper R., EML-OUNL (Open University of the Netherlands' Educational Modeling Language), Modeling Units of Study from a Pedagogical Perspective, (2001).

Le Maître, J., Bruno, E., Murisasco, E., «Temporalisation d'un document multimédia», *Document Numérique,* vol. 8, n° 4, p. 125-141, (2004).

Lebart L., Morineau A., Piron M., « Statistique exploratoire multidimensionnelle », 2e cycle, 2e ´edition, Editions Dunod (1997).

Lebrun M., « Des technologies pour enseigner et apprendre », *De Boeck* (2ème édition), (2002).

Lenne D., Abel M.-H., Moulin C. et Benayache A., « Mémoire de formation et apprentissage ». *EIAH 2005*, Montpellier, (2005).

Lin, D., « An information-theoric definition of similarity », *In Proceedings of the 15th international conference on Machine Learning*, p. 296-304, (1998).

Lim, L., Wang, M., Padmanabhan, S., Vitter, J.S., Agarwal, R.C.: « Efficient update indexes for dymnamically changing web documents ». Published online: 2 March 2007., *http://www.cs.duke.edu/~jsv/Papers/LWP05.landmarkdiff.pdf*. Springer Science + Business Media, LLC (2007).

LOM standard, document IEEE 1484.12.1-2002, (2002).

Maedche A. et Staab S., «Ontology Learning». *Handbook on Ontologies*, S. Staab, R. Stubers (Eds.), p. 173-190, (2004).

Manber, U., Wu, S.: GLIMPSE: « a tool to search through entire file systems ». *the Winter 1994 USENIX Conf.*, p. 23–32. (1994).

McBride, B., Jena: « Implementing the rdf model and syntax specification », http://wwwuk.pl.hp.com/people/bwm/papers/20001221-paper/, Hewlett Packard Laboratories, (2001).

Miller, L., Seaborne, A., Reggiori, A. « Three implementations of squishql, a simple rdf query language », *In Proceedings of the International Semantic Web Conference*, p. 423-435, (2002).

Moldovan D., Harabagiu S., Pasca M., Mihal-cea R., Goodrum R., Girju R., Rus V., « LASSO: A tool for surfing the answer net ». *Proceedings of the 8th Text Retrieval Conference (TREU-8),* (1999).

Mothe, J., « Modèle connexionniste pour la recherché d'informations : Exepansion dirigée de requêtes et apprentissage », *Thèse de l'Université Paul Sabatier* (1994).

Page, L., Brin, S.: « The anatomy of a large-scale hypertextual web search engine ». *Proceedings of the 7th Intl.WWW Conf.*, p. 107–117 (1998).

Pogodalla, V. L. et Dury J. Y, « Réflexions sur la modélisation des Documents », *Information – Interaction - Intelligence, Cépaduès*, Volume 4, n°1, p. 19-38, (2004) http://www.revue-i3.org/volume04/numero01/revue_i3_04_01_02.pdf.

Ponte, J., M., Croft, W. B. «A language modeling approach to information retrieval ». *21st annual international ACM SIGIR conference on Research and Development in Information Retrieval*, p. 275–281, Melbourne, Australia. (1998).

Porter M., « An algorithm for suffix stripping ». *Program*, 14(3):130–137. (1980)

Psyché V., Bourdeau J., Nkambou R. et Mizoguchi R. «Making Learning Design Standards Work with an Ontology of Educational Theories». *AIED* (2005).

Quillian, M.R., «Semantic Memory», M. Minsky (Ed.), *Semantic Information Processing*, M.I.T. Press, Cambridge, (1968).

Rada, R., Mili, H., Bicknell, E. et Blettner, M., « Development and application of a metric on semantic nets». *Systems, Man and Cybernetics*, IEEE Transactions on, 19(1): p. 17-30. (1989).

Raimond, Y., Abdallah, S., Sandler, M., Giasson, F. « The music ontology », *8th International Conference on Music Information Retrieval*, ISMIR, Vienna, Austria, (2007).

Resnik, P. « Using information content to evaluate semantic similarity in a taxonomy ». *IJCAI*, p. 448-453, (1995).

Reymonet, A., Thomas, J., Aussenac-Gilles, N.. «Ontology Based Information Retrieval: an application to automotive diagnosis». *International Workshop on Principles of Diagnosis (DX 2009)*, Stockholm, Mattias Nyberg, Erik Frisk, Mathias Krisander, Jan Aslund (Eds.), Linköping University, Institut of Technology, p. 9-14, juin (2009).

Robertson S. E., Sparck Jones K., « Relevance weighting of search terms ». *Journal of the American Society for Information Sciences*, 27 (3), p 129-146, (1976).

Robertson S. E., «The probabilistic character of relevance. » Inf. Process. Manage. 13(4), p. 247-251, (1977).

Rocha, C., Schwabe, D., de Aragão, M.P. « A Hybrid Approach for Searching in the Semantic Web », *In Proceedings of the 13th International World Wide Web Conference*, p. 374-383, (2004).

Roisin, C., «Authoring Structured Multimedia documents», *25th Conference on Current Trends in Theory and Practice of Informatics (SOFSEM'98)*, édité par B. Rovan, p. 222-239, Springer, LNCS 1521, Jasna, Slovakia, (1998).

Roisin C. et Sèdes, F., «Introduction », *Document numérique* 2004/4, vol. 8, p. 7-9, (2004).

Rousseaux, F., «Une contribution de l'intelligence artificielle et de l'apprentissage symbolique automatique à l'élaboration d'un modèle d'enseignement de l'écoute musicale», *Thèse de l'université de Paris 6* (1990).

Rousseaux, F, Bonardi, A, «Parcours et constituer nos collections numériques»,*10ème Colloque International sur le Document Electronique (CIDE 10)*, Nancy, (2007).

Salton, G., « The Smart Retrieval System », *Prentice Hall, Englewood Cliffs*, NJ, (1971)

Salton, G., Fox, E.A., Wu., H., « Extended Boolean information retrieval system». CACM 26(11), pp. 1022-1036, (1983).

Salton, G., Allan, J., Buckley, C.: « Approaches to passage retrieval in full text information systems ». *Korfhage, R., Rasmussen, E.M., Willett, P. (eds.) Proceedings of the 16th Annual. Intl. ACM-SIGIR Conf*, p. 49–58. (1993).

Seco, N., Veale, T. et Hayes, J. «An intrinsic information content metric for semantic similarity in Wordnet». *Proceedings of ECAI'2004, the 16th European Conference on Artificial Intelligence.* (2004).

Song, J. F., Ming Z. W., Dong X. W., Hui L. G., Ning X. Z., « Ontology-based Information Retrieval Model for the Semantic Web ». *International Conference on e-Technology, e-Commerce and e-Service, EEE '05, p 152 - 155,* (2005).

Spärck Jones, K, «A statistical interpretation of term specificity and its application in retrieval», *Journal of Documentation 28 (1)*: 11–21 (1972).

Stern, Y. « Les quatre dimensions du document ». *Document numérique,* vol.1 n°1, p. 55-60, (1997).

Studer R., Benjamins, V., R. et Fensel, D. «Knowledge Engineering: Principles and Methods. *Data and Knowledge Engineering* (DKE)», 25(1-2), p. 161-197, (1998).

Tomasic, A., Garcìa-Molina, H., Shoens K. «Incremental Updates of Inverted Lists for Text Document Retrieval». *Proceedings of the 1994 ACM SIGMOD.* Conference, pages 289–300, New York, USA, ACM Press. (1994).

Tomassen S. L., Gulla J. A., Strasunskas D., « Document Space Adapted Ontology: Application in Query Enrichment ». *11th International Conference on Applications of Natural Language toInformation Systems*. Springer, Klagenfurt, Austria, (2006).

Ukkonen, E.: « Algorithms for approximate string matching ». *Inf. Control 64*, p. 100-118, (1985).

Vallet, D., Fernández, M., Castells, P, « An Ontology-Based Information Retrieval Model », *In Proceedings of the 2nd European Semantic Web Conference*, p. 455-470, (2005).

Van Rijsbergen C. J., « Information Retrieval ». *Butterworths*, London, 2nd edition, (1979).

Vidal P., Broisin J., Duval E, Ternier S. Normalisation et stanardisation des objets d'apprentissages : l'expérience ARIADNE. *Colloque « miage et e-mi@ge », Marrakech, ESG, p. 48-64,* (2004).

Voorhees E.M., Query expansion using lexical-semantic relations, *Proceedings of the 17th International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 61-69, (1994).

Witten, I., H., Timothy, C., B., Moffat, A., « Managing Gigabytes: Compressing and Indexing Documents and Images », John Wiley & Sons, Inc., New York, NY, (1994).

Wu Z., Palmer M., «Verb semantics and lexical selection», *Proceedings of the 23rd Annual Meetings of the Associations for Computational Linguistics*, p. 133-138, (1994).

Xiaomeng S., Atle J. G., « An information retrieval approach to ontology mapping ». *Data & Knowledge Engineering*, Vol. 58 Issue 1, p. 47-69, (2006).

Zhao Y., Zhang J., Guan B., Hu J., Wang W., « The Development of Intelligent Retrieval Algorithm Ontology-based And Its Application in Bearing production Information System ». *11th International Conference on Computer Supported Cooperative Work in Design, 2007. CSCWD'07,* p. 722 – 727, (2007).

Adobe, Portable Document Format (PDF), http://partners.adobe.com/asn/developer/acrosdk/docs.html#filefmtspecs .

Adobe, Postscript Reference Manual, http://partners.adobe.com:80/asn/developer/pdfs/tn/psrefman.pdf .

AFNOR : Vocabulaire de la documentation (2ème éd. - 1987). (Les dossiers de la normalisation, ISSN 0297-4827) ISBN 2-12-484221-8 :

CREPUQ – NOVASYS (2003). *La description normalisée des ressources. Vers un patrimoine éducatif.* Montréal : Québec.

E-TUD, http://www.e-tud.com/encyclo_e-learning.htm

IMSLD, http://www.imsglobal.org/learningdesign/, (2003).

MIDI Specification, http://www.midi.org/about-midi/specshome.shtml

MusicXML Definition, Michael Good (Recordare), http://www.musicxml.org/xml.html (2002)

NIFF, http://www.music-notation.info/en/niffml/niffml.html

Play Code, http://www.dodiesis.com

SCORM, Le modèle SCORM (2004). http://www.adlnet.org

W3C Consortium (2004), OWL Specification Development, http://www.w3.org/2004/OWL/#specs , (2004).

PUBLICATIONS DE L'AUTEUR DE CETTE THESE

Le groupe auquel j'appartiens a fait le choix d'ordonner alphabétiquement les noms des auteurs des publications.

Articles de revues internationales avec actes édités et comité de lecture

- Benoit Encelle, Nadine Jessel, Josiane Mothe, Bachelin Jhonn Victorino Ralalason. BMML: Braille Music Markup Language. Dans : Open Information Science Journal, Bentham Science Publishers, Vol. 3, p. 123-135, août 2009.
Accès : http://bentham.org/open/toisj/openaccess2.htm

- Nathalie Hernandez, Josiane Mothe, Bachelin Johnn Victorino Ralalason, Andriantiana Bertin Ramamonjisoa , Patricia Stolf. A Model to Represent the Facets of Learning Objects. Dans : Interdisciplinary Journal of E-Learning and Learning Objects, Informing Science Institute, Santa Rosa - USA, Vol. 4, p. 65-82, janvier 2008.
Accès : ftp://ftp.irit.fr/IRIT/SIG/2008_IJKLO_HMRRS.pdf

Articles de revues nationales avec actes édités et comité de lecture

- Nathalie Hernandez, Josiane Mothe, Andriantiana Bertin Ramamonjisoa, Bachelin Johnn Victorino Ralalason, Patricia Stolf. Indexation multi-facette des ressources pédagogiques pour faciliter leur ré-utilisation. Dans : Revue des Nouvelles Technologies de l'Information, Cépaduès Editions, Numéro spécial Modelisation des Connaissances, 2008.

Conférences et workshops internationaux avec actes édités et comité de lecture

- Benoit Encelle, Nadine Jessel, Josiane Mothe, Bachelin J. V. Ralalason. BMML: Braille Music Markup Language. Dans: The Sixth International Workshop on XML Technology and Applications (XMLTech'08). Monte Carlo Resort, Las Vegas, Nevada, USA, 2008.

- Nathalie Hernandez, Josiane Mothe, Bachelin J. V. Ralalason, Andriantiana Bertin Ramamonjisoa , Patricia Stolf. Multi-facet indexing for learning objects reuse. Dans : Computer Science & Information Technology Education, Pointe-aux-Sables, 16-NOV-07-18-NOV-07, Institute for Scientific Information (ISI), p. 309-322, novembre 2007.
Accès : http://www.irit.fr/~Josiane.Mothe/pub/CSITEd07.pdf

- Nathalie Hernandez, Josiane Mothe, Bachelin J. V. Ralalason, Patricia Stolf. Modèle de représentation sémantique des documents électroniques pour leur réutilisabilité dans l'apprentissage en ligne. Dans : Colloque International sur le Document Électronique (CIDE 2006), Fribourg (Suisse), 18-SEP-06-20-SEP-06, EUROPIA, p. 181-198, 2006.

Conférences et workshops nationaux avec actes édités et comité de lecture

•Damien Dudognon, Bachelin J. V. Ralalason, *ProxiGénéa : Une mesure de similarité conceptuelle*. Dans : Veille Stratégique Scientifique et Technologique (VSST 2010), Université Paul Sabatier Toulouse, 25/10/2010-29/10/2010, UPC-IRIT-SFBA.

•Gilles Hubert, Josiane Mothe, Bachelin J. V. Ralalason, Andriantiana Bertin Ramamonjisoa . Modèle d'indexation dynamique à base d'ontologies. Dans : Conférence francophone en Recherche d'Information et Applications (CORIA 2009), Belambra de la Presqu'île de Giens, dans le Var, 05/05/2009-07/05/2009, Laboratoire des Sciences de l'Information et des Systèmes (LSIS), p. 169-184, 2009.

Rapports

•Philippe Laublet, Nathalie Aussenac-Gilles, Valérie Camps, Pierre Glize, Nathalie Hernandez, H. Maurel, Mohamed Mbarki, Josiane Mothe, Bachelin Jhonn Victorino Ralalason, Axel Reymonet, Bernard Rothenburger, Zied Sellami, Jérôme Thomas, Anis Tissaoui. *Projet ANR 07 TLOG 004 01 - DYNAMO (DYNAMic Ontology for information retrieval) : Etat de l'art - Livrable lot 2*. Rapport de contrat, Dynamo 2.1, IRIT, décembre 2009. Accès : ftp://ftp.irit.fr/IRIT/IC3/Rapport_lot2_integre_F7.pdf

•Nathalie Hernandez, Gilles Hubert, Josiane Mothe, Bachelin Jhonn Victorino Ralalason. *RI et Ontologies – Etat de l'art 2008*. Rapport de recherche, IRIT/RR—2008-14--FR, IRIT, juillet. 2008. Accès : ftp://ftp.irit.fr/IRIT/SIG/2008_RA-14-FR_HHMR.pdf

•Nadine Jessel, Bachelin Ralalason, Enrico Bortolazzi, « *D3.1 BMML Markup Language* », Preservation and unification of new and existing Braille Music digital sources for a new access methodology, *Contrapunctus.*
Accès: http://www.punctus.org/public/documents/D3_1.pdf

LISTE DES FIGURES :

LISTE DES TABLES :

ANNEXES

Les standards LOM, SCORM et IMS-LD :

Annexe A : LOM

1. Généralité : caractéristiques indépendantes du contexte comme identifiant (un identificateur global unique) ou titre (le nom de la ressource) ou langage (la langue utilisée principalement par la ressource pour communiquer avec l'utilisateur), catalogue…
2. Cycle de vie : caractéristiques relatives au cycle de vie, comme Version ou Etat (Brouillon, Final, Révisé, Non disponible).
3. Méta-métadonnées : caractéristiques de la description elle-même comme Identifiant, Contribution (personnes ayant participé à l'élaboration des métadonnées), Catalogue, langage…
4. Technique : caractéristiques techniques comme le format (du logiciel nécessaire pour accéder à la ressource), la taille…
5. Pédagogie : caractéristiques pédagogiques.

   (a) *Type d'interactivité* : le type d'interaction entre la ressource et l'utilisateur typique *(Actif, Descriptif, Combiné)* ;

   (b) *Type de ressource pédagogique*: le type pédagogique *(Exercice, Simulation, Questionnaire, Diagramme, Figure, Graphique, Index, Diapositive, Tableau, Texte narratif, Examen, Expérience, Enoncé d'un problème, Auto-évaluation, Exposé)*, peut être présent plusieurs fois ;

   (c) *Niveau d'interactivité* : degré d'interactivité (très faible, faible, moyen, élevé, très élevé) ;

   (d) *Densité sémantique* : (très faible, faible, moyen, élevé, très élevé)

   (e) *Rôle présumé de l'utilisateur final* : utilisateur de la ressource (Auteur, Enseignant, Apprenant, Gestionnaire) ;

   (f) *Contexte* : environnement d'utilisation de la ressource (Ecole primaire/secondaire, post-secondaire, formation, autre…) ;

   (g) *Tranche d'âge*: âge de l'utilisateur ;

   (h) *Difficulté* : difficulté de la ressource (Très facile, Facile, Moyen, Difficile, Très difficile) ;

   (i) *Temps d'apprentissage moyen* : temps approximatif ou typique pour travailler avec la ressource ;

   (j) *Description* : commentaires sur l'utilisation de la ressource ;

   (k) *Langage* : la langue de l'utilisateur.
6. Droits : Coûts, copyrights, description…
7. Relation : caractéristiques exprimant les liens avec d'autres ressources comme Type (nature de la relation).
8. Annotation : commentaires sur l'utilisation pédagogique de la ressource : Auteur, date, description.
9. Classification : caractéristiques de la ressource décrites par des entrées dans des systèmes de classification : but, classification de références, chemin…

Annexe B : Profil d'application

Le standard LOM fournit une base concrète de départ pour la normalisation et l'indexation des ressources d'enseignement utilisées dans les systèmes de gestion de la formation. Cependant, il y a des critiques qui lui sont faites par la plupart des acteurs de la normalisation (Bourda, 2001) (De La Passardière , 2004) dont :

Une certaine incohérence entre la définition générique des objets pédagogiques proposée par IEEE et les éléments permettant de les décrire (la prise en compte d'entités non numériques), l'intégration au sein d'un même modèle des entités de niveaux conceptuellement très différents : les ressources nécessaires à la mise en place d'activités pédagogiques et les activités elles-mêmes.

- le fait que l'unité d'indexation soit le fichier, qui représente une unité technique et non pédagogique,
- le fait qu'un cours complet soit indexé au même titre qu'un unique exercice ou qu'une image,
- le fait que le problème de la signification des termes choisis et de la définition des métadonnées ne soit pas complètement résolu,
- les ambiguïtés du modèle restent un frein à un réel usage pratique.

Dans un profil d'application, les éléments obligatoires assurent un minimum d'information pour une ressource donnée tandis que les éléments à caractère optionnel simplifient le travail d'indexation (c'est-à-dire la mise en conformité des contenus) et affectent aussi la cohérence de l'ensemble.

Ainsi, des définitions de profil d'application sont proposées :

«Un profil d'application est un assemblage d'éléments de métadonnées choisis à partir d'un ou plusieurs schémas de métadonnées et combiné à un schéma composé […] Son objet est d'adapter des schémas existants pour constituer un ensemble taillé à la mesure des exigences fonctionnelles d'une application particulière, tout en restant inter opérable avec les schémas d'origine. » (Duval et al., 2002).

«Un profil d'application est une sélection d'éléments d'une norme, d'un standard ou d'une spécification formant ainsi un sous-ensemble adapté aux besoins des groupes qui l'utilisent. Le sous-ensemble d'éléments est défini pour fournir un cadre d'opération» (CREPUQ, 2003).

Il s'agit donc d'adapter les standards pour répondre aux besoins spécifiques et concrets des utilisateurs. De fait, cela signifie interpréter, raffiner, étendre ou parfois même simplifier les syntaxes et les sémantiques. Ce travail d'adaptation conduit à la mise en place d'un «profil d'application».

Dans notre cas, le profil d'application permet donc de définir, pour une application donnée, ou une formation donnée quelles sont les métadonnées (issues d'un ou plusieurs schémas) qui ont un intérêt pour cette application.

Voici des exemples de profil d'application de LOM :

- CanCore (Canadian Core Learning Resource Metadata Application Profile) propose 61 éléments optionnels,
- le profil finlandais réalisé en Finlande en 2003 suggère 13 éléments obligatoires et 10 recommandés,
- le profil The Gateway to Educational Materials (GEM 2.0),

- Celebrate du projet européen Celebrate comprend 20 éléments obligatoires, 12 recommandés et 30 facultatifs,
- ManUel, créé en France en 2003, propose 11 éléments obligatoires, 15 recommandés et 21 facultatifs,
- le profil LOM-FR, établi en 2005 en France présente 4 éléments obligatoires, 15 recommandés et 61 facultatifs,
- le profil Licef, établi en Québec en 2003 offre 62 métadonnées facultatifs,
- le profil UK LOM Core par contre exige 22 métadonnées obligatoires.

Nous constatons que le nombre d'éléments utilisés varie d'un profil à un autre.

Annexe C : SCORM

SCORM traite les éléments suivants :

- Packaging : il a pour objectif la transmission d'un contenu d'une plate-forme vers une autre, l'importation ou l'exportation de contenus d'objets pédagogiques pour les mettre à disposition d'autres. Il s'intéresse également à la structuration des objets pédagogiques (Content structure),
- Métadonnées : elles sont issues de LOM et ont pour objectif de partager les informations standards qui décrivent la nature et l'objectif du contenu. Ces informations peuvent être utilisées soit pour la recherche et la découverte de documents soit pour la gestion des droits et des besoins techniques,
- Communication ou environnement d'exécution : détermine la communication avec un environnement web. La notion d'environnement est également présente dans IMS-LD,
- Séquencement et navigation : définit une méthode de représentation de la navigation entre objets d'apprentissages. Spécifiquement, il décrit les branchements et le flux d'activités d'apprentissages en termes d'arbre d'activité,
- Agrégation de contenu : elle distingue trois niveaux de ressources :
    - La *ressource numérique élémentaire* (asset) constitue la brique élémentaire : Il peut s'agir d'un document simple (image JPEG ou GIF, son WAV ou MP3, page Web) mais également de tout ensemble d'informations pouvant être délivré vers un client Web (document Flash, code Javascript, applet Java, etc.).
    - Un *objet de contenu partageable* (SCO) est un ensemble cohérent de ressources numériques élémentaires. Il peut être contrôlé depuis une plate-forme d'apprentissage en ligne (LMS). Respectant le protocole d'exécution SCORM, il représente le plus bas niveau de granularité pouvant faire l'objet d'un suivi.
    - Un *agrégat de contenu* (Content Aggregation) est un ensemble de ressources pédagogiques structurées de façon cohérente au sein d'une entité de plus haut niveau, telle qu'un cours, un chapitre, un module, etc.

Extrait du fichier imsmanifest.xml

Tout package d'objet pédagogique comprend le fichier imsmanifest.xml qui décrit la structure de l'objet pédagogique et de ses composants. Il contient aussi les différentes métadonnées décrivant l'objet pédagogique en question.

```
<manifest identifier="MANIFEST-F374AE0A-C32F-3A51-AD8B-6ED3E5962CAD" xsi:schemaLocation="http://www.imsglobal.org/xsd/imscp_v1
imscp_v1p1.xsd http://www.imsglobal.org/xsd/imsmd_v1p2 imsmd_v1p2p2.xsd http://www.adlnet.org/xsd/adlcp_rootv1p2 adlcp_rootv1p2.xsd">
 - <metadata>
     <schema>ADL SCORM</schema>
     <schemaversion>1.2</schemaversion>
     <adlcp:location/>
   - <imsmd:lom>
     - <imsmd:general>
         <imsmd:identifier>B210ALGANIMUNIONC01</imsmd:identifier>
       - <imsmd:title>
           <imsmd:langstring xml:lang="en">Animation union</imsmd:langstring>
         </imsmd:title>
         <imsmd:language>fr</imsmd:language>
       - <imsmd:keyword>
           <imsmd:langstring xml:lang="en">Algèbre relationnelle, Union</imsmd:langstring>
         </imsmd:keyword>
       - <imsmd:structure>
         - <imsmd:source>
             <imsmd:langstring xml:lang="en"/>
           </imsmd:source>
         - <imsmd:value>
             <imsmd:langstring xml:lang="x-none">Atomic</imsmd:langstring>
           </imsmd:value>
         </imsmd:structure>
       </imsmd:general>
     - <imsmd:lifecycle>
       - <imsmd:version>
           <imsmd:langstring xml:lang="en">1.0</imsmd:langstring>
         </imsmd:version>
```

Les métadonnées qui se trouvent dans le fichier imsmanifest.xml sont issues du profil d'application EMIAGE comme illustré dans la Figure 25.

Annexe D : IMS-LD

Modèle Conceptuel d'IMS-LD au niveau C : [IMSLD03]

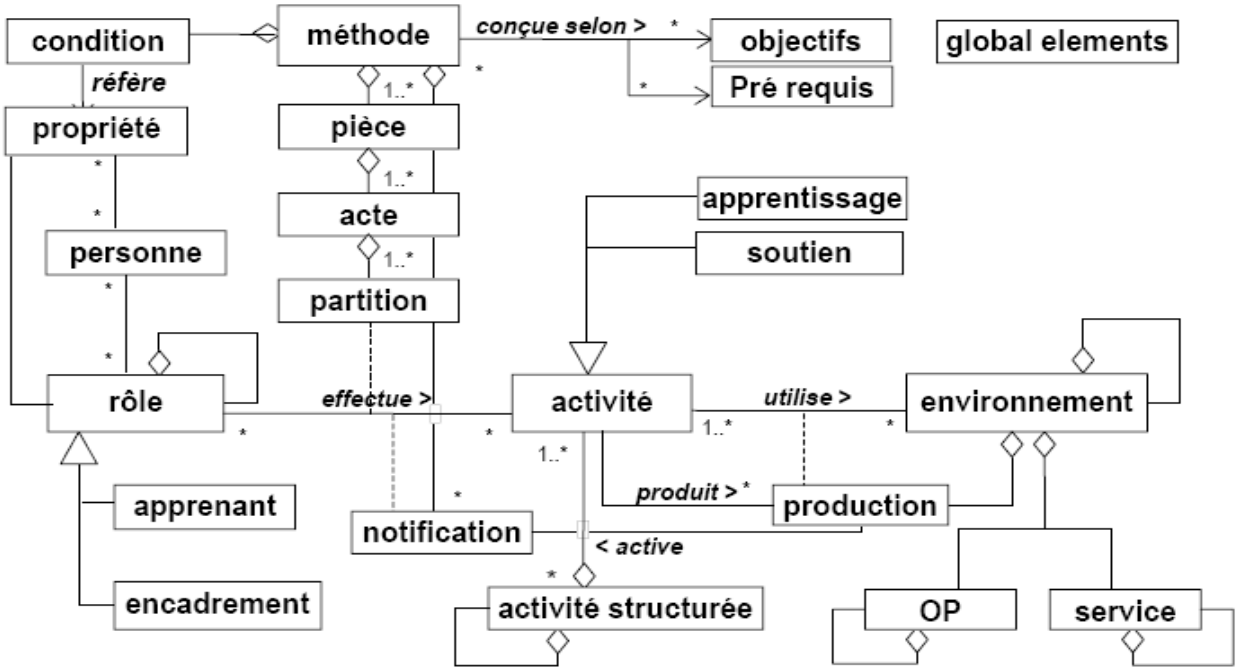

Figure 57 : Modèle conceptuel d'IMS-LD au niveau C.

Objectifs pédagogiques :

Ce sont des objectifs à atteindre à la fin d'une unité d'apprentissage. Ils peuvent être spécifiés soit au niveau du scénario soit au niveau des activités.

Prérequis :

Ils correspondent aux conditions que doit remplir l'apprenant pour pouvoir suivre l'unité d'enseignement. On peut définir les prérequis au niveau du scénario et au niveau des activités.

Propriétés :

La notion de propriété n'apparaît qu'aux niveaux B et C. Elle permet la constitution d'un dossier de l'apprenant ou du rôle. C'est un élément essentiel pour la réalisation d'unités d'apprentissage personnalisables.

Éléments Globaux :

C'est une fonction externe au Learning Design. Ces éléments sont représentés dans le Learning Design parce qu'ils sont indispensables aux scénarios qui les utilisent.

Activité pédagogique :

Learning Activity est composé d'une activity-description et de quelques autres éléments.

Support Activité :

C'est une activité de soutien, elle contient les mêmes éléments qu'une activité d'apprentissage excepté qu'elle n'a pas d'objectif d'apprentissage.

Structure-Activité :

Une activity-structure est une agrégation d'activités (d'apprentissage ou d'assistance).

Services :

Les activités d'apprentissages nécessitent des services communs qui ne peuvent être traités comme les objets pédagogiques. Il s’agit par exemple des forums, des services de mails ou une recherche documentaire (donc il s’agit d’outils).

Conditions :

Cette section n'existe qu'au niveau B et C. Elle est basée sur les valeurs prises par les propriétés d'un dossier précis.

Notification :

Cette section n'existe qu'au niveau C. Elle permet d'envoyer un message à un élément du Learning Design afin de déclencher une réaction prévue.